numero 145
settembre 1980
anno XIII
volume XXV

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Divieto alla produzione di materiale fissile per armi

Una decisione di questo genere raggiungerebbe due scopi: interrompere l'ulteriore accumulo di armi nucleari da parte degli stati che già le detengono e impedirne la diffusione ai paesi che non ne sono ancora dotati

di William Epstein

La sospensione della procedura di ratifica del trattato SALT 2 da parte del senato degli Stati Uniti in seguito all'intervento militare dell'Unione Sovietica in Afghanistan non rappresenta certo la fine degli sforzi internazionali intesi a rallentare la corsa agli armamenti nucleari fra le due superpotenze. Oltre al procedimento bilaterale dei SALT, sono in corso o in progetto vari negoziati multilaterali per il controllo degli armamenti e l'amministrazione Carter ha promesso solennemente la partecipazione continua degli USA a questi tentativi. Attualmente, per esempio, sono in corso colloqui fra i rappresentanti degli USA, dell'URSS e del Regno Unito per un trattato generale e completo che interdica tutte le esplosioni nucleari sperimentali. Inoltre, è in programma entro quest'anno la seconda conferenza quinquennale allo scopo di riesaminare il funzionamento del Trattato di non proliferazione.

Nel comunicato congiunto emesso l'anno scorso dagli USA e dall'URSS al momento della firma del Trattato SALT 2, tanto la messa al bando totale degli esperimenti nucleari quanto il Trattato di non proliferazione figuravano al primo posto nell'elenco delle questioni importanti relative alla limitazione degli armamenti. In quel documento si affermava che erano stati fatti «sicuri progressi» nei negoziati per la messa al bando totale degli esperimenti e si confermavano le intenzioni delle due superpotenze di lavorare con il Regno Unito «per completare il più presto possibile la preparazione di tale trattato». Quanto al Trattato di non proliferazione, il presidente americano Carter e il presidente sovietico Breznev si limitavano a mettere in rilievo «la profonda minaccia posta alla sicurezza mondiale dalla proliferazione delle armi nucleari» e si dichiaravano d'accordo sul fatto che gli stati che detengono armi nucleari «hanno una particolare responsabilità nel mostrare moderazione» a questo riguardo: essi inoltre «affermavano la loro comune convinzione che sono necessari ulteriori sforzi» per rafforzare il regime di non proliferazione.

I due capi di stato avevano chiaramente ragione a mettere in rilievo la necessità di ulteriori sforzi, se non altro per assicurare il successo della Conferenza di revisione del Trattato di non proliferazione. Gli stati non detentori di armi nucleari non si può certo dire che siano stati colpiti favorevolmente dai risultati dell'accordo SALT 2, che, se pure venisse ratificato, non fermerebbe la produzione di una sola nuova arma strategica attualmente in fase di sviluppo e permetterebbe a ognuna delle due superpotenze di aggiungere migliaia di nuove testate nucleari strategiche ai loro arsenali già eccessivi. Quasi tutti gli stati non detentori di armi nucleari danno senza dubbio l'impressione di considerare l'accordo SALT 2 come un passo avanti, ma certamente non pensano che esso si conformi adeguatamente all'articolo VI del Trattato di non proliferazione, il quale impone ai tre stati firmatari detentori di armi nucleari (USA, URSS e Regno Unito) la «cessazione entro breve tempo della corsa agli armamenti nucleari».

Che cosa si può fare, in queste circostanze, per porre un freno tanto alla proliferazione «verticale» (all'aumento cioè del numero e del tipo di armi nucleari in possesso degli stati che già ne sono detentori), quanto alla proliferazione «orizzontale» (vale a dire all'ulteriore diffusione delle armi nucleari a nazioni che ancora non ne sono dotate)? Secondo la mia opinione, basata su più di un trentennio di lavoro incentrato su questioni internazionali relative al controllo degli armamenti, due sono in questo momento le misure che potrebbero avere un effetto salutare non solo sulla rivalità fra le due superpotenze in fatto di armamenti strategici, ma anche sull'atteggiamento degli stati non dotati di armi nucleari nei confronti di tutto il regime di non proliferazione. Una sarebbe il sollecito conseguimento di un Trattato per la messa al bando totale degli esperimenti. L'altra sarebbe l'assunzione dell'impegno da parte di tutti gli stati interessati - detentori e non detentori di armi nucleari - di avviare rapidamente dei negoziati per un trattato che metta al bando la produzione di materiale fissile per armi nucleari.

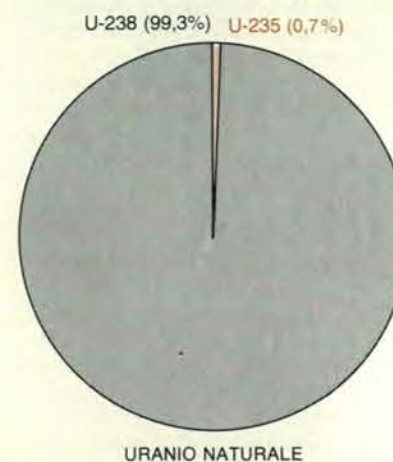
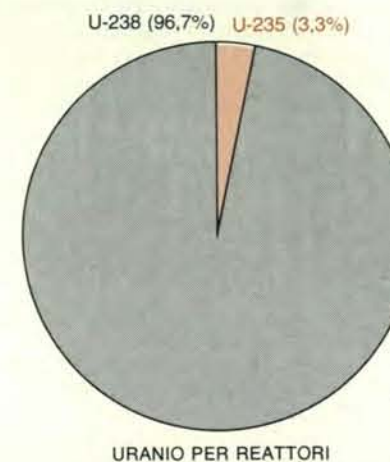
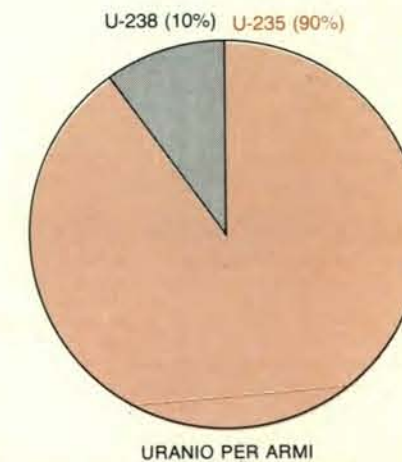
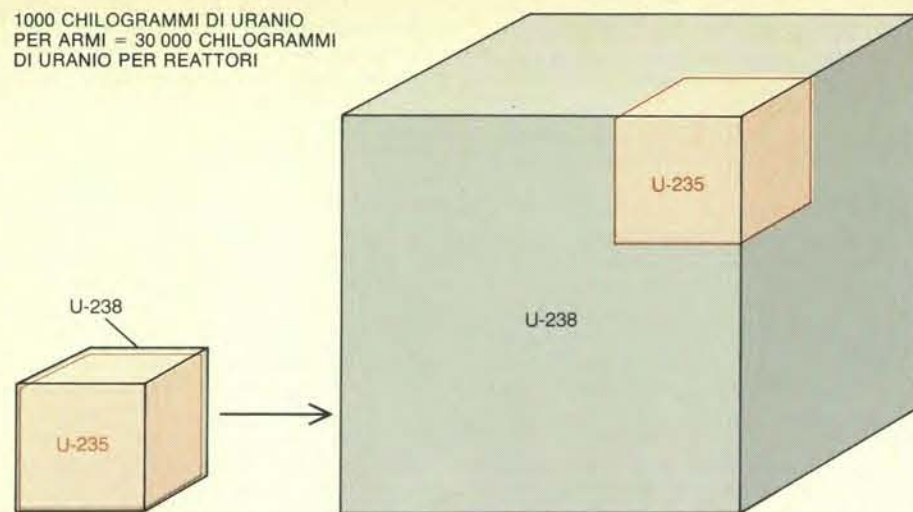
Non è certo necessario discutere in questa sede i particolari dell'imminente interdizione totale degli esperimenti. Dal 1963, quando fu concluso il Trattato per la messa al bando parziale degli esperimenti (che proibiva le esplosioni nucleari nell'atmosfera, sott'acqua e nello spazio), un'interdizione totale corrisponde in pratica alla messa al bando degli esperimenti sotterranei. Nella storia del controllo degli armamenti e del disarmo l'interdizione totale degli esperimenti è stata oggetto di discussioni e di negoziati più di qualsiasi altra questione. Nel testo del Trattato per la messa al bando parziale degli esperimenti e ancora nel Trattato di non proliferazione è contenuta la promessa dei tre stati firmatari detentori di armi nucleari di cercare «per sempre la cessazione di tutte le esplosioni sperimentali». Per gli stati non detentori di armi nucleari il conseguimento di un'interdizione totale degli esperimenti è una prova decisiva della serietà delle intenzioni degli stati detentori di fermare la corsa agli armamenti nucleari: fu questa, in effetti, la loro principale richiesta alla prima Conferenza di revisione del trattato di non proliferazione, che si tenne a Ginevra nel 1975. In realtà, se nel 1963 fossero riusciti a raggiungere un accordo per l'interdizione totale degli esperimenti, i tre stati detentori di armi nucleari molto probabilmente avrebbero evitato l'impiccio della proliferazione in cui si trovano attualmente.

Forse, se si riuscisse a raggiungere entro breve tempo un accordo per l'interdizione totale degli esperimenti, sarebbe ancora possibile porre rimedio ai danni causati allora. Si dice che le tre parti impegnate nelle trattative siano molto vicine all'accordo e che si aspettavano di concludere felicemente i loro negoziati poco dopo la ratifica del SALT 2. La mancata ratifica di tale trattato potrebbe avere soltanto un effetto negativo sulle prospettive della non proliferazione e della messa al bando totale degli esperimenti.

A giudicare dalle apparenze, sembrerebbe che il conseguimento della messa al bando totale degli esperimenti sia un passo relativamente facile, dal momento che quasi tutti i vantaggi di carattere militare che probabilmente si potrebbero trarre da un gran numero di esperimenti di armi nucleari fanno già parte integrante degli arsenali nucleari delle due superpotenze. Certo, non è da escludere che ad alcuni funzionari delle forze armate e ad alcuni addetti alla progettazione di armi nucleari di entrambe le parti non farebbe piacere veder chiudere la porta a ogni ulteriore sperimentazione di armi nucleari, in quanto essi presumibilmente preferirebbero continuare a perfezionare e a migliorare la grande varietà di armi nucleari sofisticate che esistono attualmente, e a svilupparne di nuove per future generazioni di vettori. Ogni possibile vantaggio derivante dalla continua sperimentazione sarebbe però irrilevante rispetto ai gravi rischi di proliferazione che una politica di questo genere presenterebbe.

A me pare che un passo altrettanto importante per porre un freno alla corsa agli armamenti sarebbe quello di mettere al bando la produzione di materiale fissile per armi nucleari. (A quasi tutti i fini pratici, ciò significa uranio naturale arricchito a circa il 90 per cento nell'isotopo uranio 235, prontamente fissile, e plutonio ottenuto nei reattori contenente una concentrazione abbastanza elevata dell'isotopo plutonio 239.) Anche questa proposta è da molti anni oggetto di discussioni, anche se su di essa non si sono mai svolti negoziati formali. In origine l'idea fu avanzata come modo per «tagliare» ossia arrestare la produzione di materiale fissile idoneo alla fabbricazione di armi nucleari in speciali attrezzature adibite alla produzione militare nell'ambito degli stati già detentori di armi nucleari. In anni recenti, però, a mano a mano che molti altri paesi acquisivano la capacità di produrre plutonio e uranio arricchito nel corso dei loro programmi nucleari a fini pacifici, è diventato più corretto parlare di una più ampia proibizione della produzione di materiale idoneo alla fabbricazione di armi sia da parte degli stati detentori di armi nucleari, sia da parte di quelli che ne sono ancora sprovvisti. Ciò nonostante, negli ambienti interessati al controllo degli armamenti il termine «taglio» *(cut off)* ha continuato a essere usato, in senso allargato, per indicare entrambe le circostanze.

La proposta messa al bando della produzione sarebbe un passo successivo logico e possibile lungo il cammino della limitazione delle armi nucleari. In effetti, i problemi inerenti al suo negoziato sarebbero piuttosto semplici, soprattutto rispetto alle particolari difficoltà e complessità di qualsiasi futuro negoziato SALT. Ancora più importante è il fatto che un trattato che prevedesse fra l'altro un'interdizione di questo genere contribuirebbe a contenere la proliferazione delle armi nucleari sia in senso verticale sia in senso orizzontale. Il concorso di molteplici circostanze ha reso oramai maturo il momento per questo passo. Esso perciò dovrebbe essere messo al primo posto nell'ordine del giorno del controllo internazionale degli armamenti, alla pari con l'interdizione totale degli esperimenti.

L'idea di un taglio della produzione di materiale fissile idoneo alla fabbricazione di armi fu avanzata per la prima volta nel discorso «Atomi per la pace» pronunciato dal presidente americano Eisenhower all'Assemblea generale delle Nazioni Unite nel 1953. La proposta di Eisenhower, che nasceva dal fallimento nel dopoguerra del piano Baruch per il

Allo stato naturale *(a sinistra)* l'uranio è costituito quasi interamente dall'isotopo uranio 238, con una piccola aggiunta dell'isotopo uranio 235, facilmente fissionabile. L'uranio leggermente arricchito, combustibile in un reattore nucleare di potenza *(al centro)*, contiene circa il 3,3 per cento di uranio 235. L'uranio molto arricchito per la fabbricazione di armi nucleari *(a destra)* contiene più del 90 per cento di uranio 235.

La conversione a usi pacifici di 1000 chilogrammi di uranio idoneo alla fabbricazione di armi nucleari, che attualmente fanno parte delle scorte militari, produrrebbe circa 30 000 chilogrammi di uranio utile per i reattori, una quantità sufficiente ad alimentare per circa un anno un reattore di potenza di 1000 megawatt del tipo ad acqua naturale. Ovviamente la quantità esatta di materiale fissile idoneo alla fabbricazione di armi nucleari che attualmente fa parte delle scorte militari dei paesi detentori di armi nucleari non è un dato di dominio pubblico. È comunque possibile farsi un'idea della quantità potenziale di combustibile nucleare che si potrebbe ricavare da questa fonte, basandosi sul fatto che nel 1964 gli USA si offrirono di trasferire a usi pacifici 60 000 chilogrammi di materiale fissile delle proprie scorte militari, se l'URSS avesse compiuto un gesto analogo, convertendone 40 000 chilogrammi. L'URSS, che disponeva di materiale fissile per la fabbricazione di armi nucleari in quantità molto inferiore a quella degli USA, respinse la proposta.

controllo internazionale dell'energia atomica, comprendeva non solo l'idea di un taglio della produzione, ma anche un appello a «contribuire, attingendo dalle scorte di uranio normale e di materiale fissile», alla creazione di un nuovo organismo internazionale che fosse dotato del potere di promuovere e regolare l'uso dell'energia atomica a scopi pacifici. (Questa parte della proposta portò all'istituzione, nel 1957, dell'International Atomic Energy Agency, o IAEA.)

I funzionari responsabili della proposta americana di «Atomi per la pace» tentavano così un nuovo modo di considerare il controllo degli armamenti nucleari e il disarmo. Il «taglio», per loro, era un modo per fissare un tetto alla quantità di materiale fissile disponibile per le armi nucleari e quindi al numero totale delle armi nucleari che si potevano fabbricare, e il trasferimento di tale materiale dalle scorte militari al nuovo organismo internazionale era un modo per incominciare a ridurre il numero totale delle armi nucleari già esistenti. Essi consideravano le due parti della proposta come misure collegate fra loro che potevano procedere in parallelo con altri sforzi tendenti a impedire l'ulteriore proliferazione delle armi nucleari sia in senso verticale sia in senso orizzontale.

Nel 1956 il segretario di stato americano Dulles, che agiva per conto degli USA, del Regno Unito, del Canada e della Francia, presentò ai cinque membri del sottocomitato della Commissione delle Nazioni Unite per il disarmo, che a quel tempo era radunato a Londra, un documento di lavoro intitolato «Proposte per misure parziali di disarmo». Tra i provvedimenti previsti dal documento di lavoro c'era la proposta che «ogni futura produzione di materiale fissile fosse usata, sotto supervisione internazionale, esclusivamente a fini non militari», e che «i firmatari si impegnassero a compiere, sotto supervisione internazionale, equi trasferimenti a scopi non militari di una parte sempre maggiore del materiale fissile precedentemente prodotto a fini militari». L'URSS respinse in blocco il documento di lavoro adducendo a pretesto che la messa al bando della produzione di materiale fissile per armi nucleari senza l'interdizione nel contempo delle armi nucleari stesse non avrebbe avuto alcun valore pratico.

L'anno dopo, nonostante l'opposizione dell'URSS, l'Assemblea generale delle Nazioni Unite approvò una mozione, basata su un abbozzo presentato dagli USA, che esortava gli stati membri a dare la priorità a un accordo sul disarmo che stabilisse numerose misure di disarmo parziale, ivi compresa «la cessazione della produzione di materiale fissile a fini militari». Era la prima volta che l'Assemblea generale approvava una mozione che riguardava specificamente il taglio della produzione.

La questione fu sollevata ancora nel 1964, quando, in un'allocuzione al Comitato delle Nazioni Unite per il disarmo, il presidente americano Johnson propose un «congelamento» della corsa agli armamenti nucleari, accompagnato da un accordo separato per un taglio della produzione di materiale fissile, che poteva incominciare con la chiusura verificata di tutti gli impianti. Più tardi, in quello stesso anno, gli USA, l'URSS e il Regno Unito annunciarono separatamente la propria decisione unilaterale di ridurre *(cut-back)* immediatamente la produzione di materiale fissile a fini militari. Gli USA spiegarono che la loro riduzione della produzione di materiale fissile, da attuarsi in un arco di tempo di quattro anni, avrebbe rappresentato (aggiunta alle precedenti riduzioni) una diminuzione complessiva del 20 per cento della produzione di plutonio e del 40 per cento della produzione di uranio arricchito. L'URSS a sua volta annunciò la propria decisione di interrompere immediatamente la costruzione di due grandi reattori destinati alla produzione di plutonio, allo scopo di ridurre sostanzialmente nel giro di pochi anni la produzione di uranio 235 per le armi nucleari e di destinare a scopi pacifici una maggiore quantità di materiale fissile. Il Regno Unito rispose facendo rilevare di avere già adottato una politica che seguiva quelle stesse linee direttrici; esso infatti aveva annunciato in precedenza che la sua produzione di uranio 235 era cessata completamente, mentre la produzione di plutonio stava finendo gradatamente. Bisogna sottolineare il fatto che nessuna delle tre parti ebbe mai a lamentarsi del fatto che una delle altre non si fosse attenuta alle proprie promesse unilaterali.

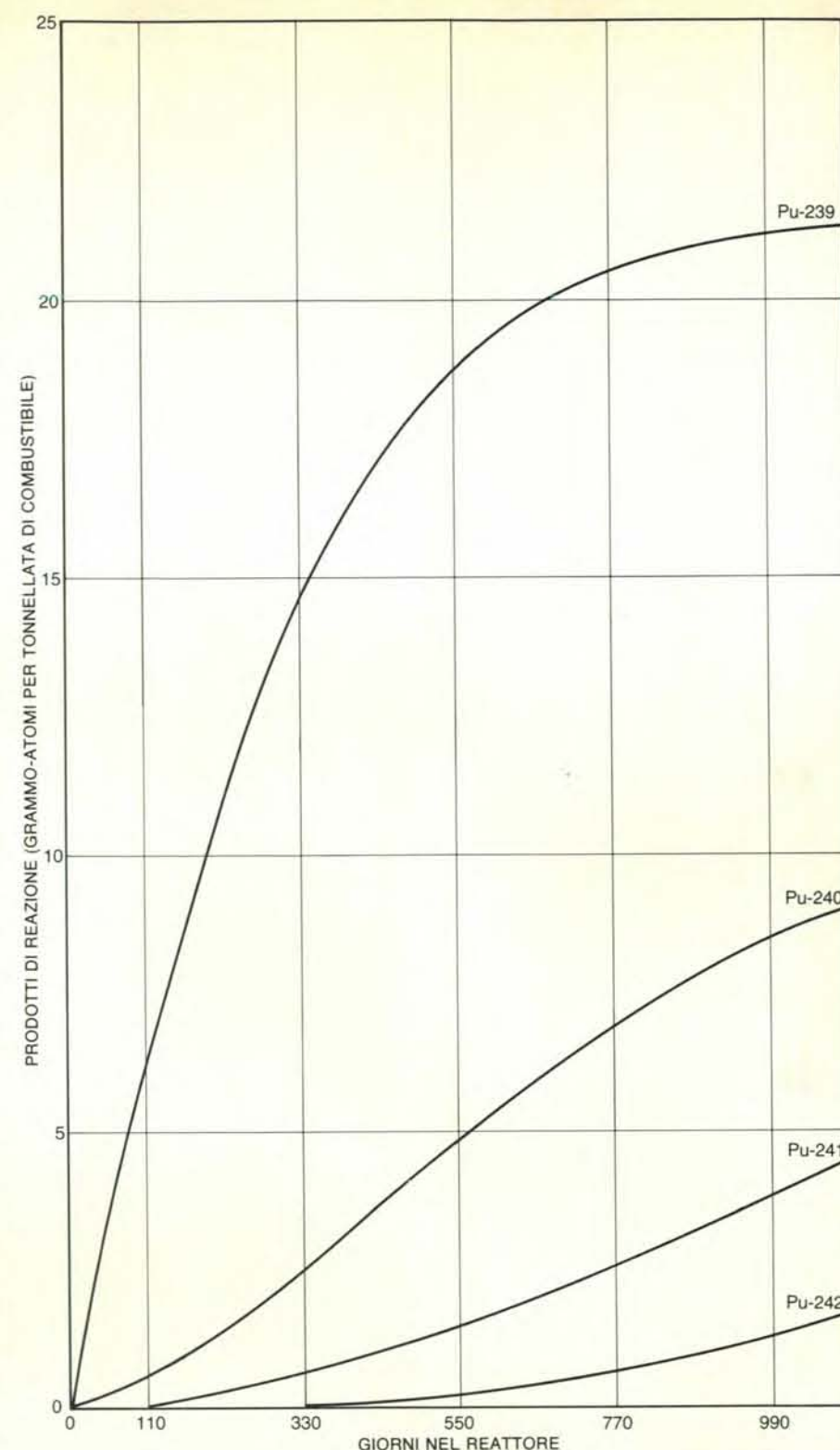
In seguito gli Stati Uniti presentarono un documento di lavoro particolareggiato, a sostegno di un taglio completo, indicando a grandi linee procedure di verifica prive di qualsiasi carattere di ingerenza e annunciando la propria intenzione di convertire a usi pacifici 60 000 chilogrammi di materiale fissile (l'URSS avrebbe dovuto convertirne 40 000 chilogrammi). Ancora una volta l'Unione Sovietica espresse i propri dubbi sulla possibilità di un accordo separato su un taglio al di fuori di un accordo generale sul disarmo.

Non desta meraviglia il fatto che negli anni cinquanta e sessanta l'URSS si sia opposta alla proposta di un taglio. Ogni potenza nucleare considera riservate tutte le informazioni relative all'entità delle proprie scorte di plutonio e di uranio arricchito fissili e continua a custodire queste informazioni come segreto militare. Il possesso di tali informazioni darebbe a un avversario una base precisa per valutare l'entità dell'arsenale nucleare di una nazione. A quel tempo per altro era ampiamente riconosciuto il fatto che, con il loro lungo vantaggio iniziale e la loro maggiore competenza tecnica, gli Stati Uniti avevano accumulato scorte di materiale fissile in misura molto ma molto maggiore dell'Unione Sovietica. Pertanto negli anni cinquanta e sessanta qualsiasi taglio della produzione, anche se accompagnato, da parte americana, da più massicci trasferimenti a scopi pacifici di materiale prelevato dalle scorte militari, avrebbe certamente operato a discapito dell'Unione Sovietica.

Gli USA però insistettero e nel 1966 presentarono tre documenti di lavoro, in cui venivano abbozzate, con maggiore abbondanza di particolari, proposte per il trasferimento di materiale fissile ottenuto dalla distruzione di armi nucleari e per l'adozione di sistemi di ispezione per controllare la chiusura degli impianti di produzione del plutonio e l'interruzione dell'attività dei reattori nucleari. La cosa però non ebbe alcun seguito nei due anni che seguirono, in quanto nel frattempo USA e URSS si erano preoccupati di negoziare un accordo sul Trattato di non proliferazione.

Nel 1969 gli USA tornarono a sottolineare l'importanza del taglio. Il presidente Nixon lo incluse, unitamente a un Trattato per la messa al bando totale degli esperimenti, fra le questioni sulle quali gli Stati Uniti avrebbero insistito agli incontri di Ginevra del Comitato per il disarmo. Per rispondere alle obiezioni dell'URSS a proposito delle norme sulla verifica proposte in precedenza, gli Stati Uniti annunciarono che nell'ambito di un accordo sul taglio non avrebbero più richiesto un'«ispezione dell'avversario», ma avrebbero fatto assegnamento sui sistemi di salvaguardie della IAEA. Evidentemente l'amministrazione Nixon era giunta alla conclusione che, anche applicato soltanto alle attrezzature «dichiarate» o a quelle chiuse, il sistema di salvaguardie della IAEA avrebbe permesso di verificare in maniera adeguata se i sovietici si attenevano o meno al trattato. Gli USA mettevano in evidenza anche il fatto che, nel contesto di un accordo per il taglio, gli stati detentori di armi nucleari avrebbero potuto accettare le stesse salvaguardie della IAEA che, in conformità al Trattato di non proliferazione, venivano imposte agli stati sprovvisti di armi nucleari. Infine, sostenevano gli USA, la tesi sovietica, secondo la quale un taglio non avrebbe avuto la men che minima utilità in quanto non toccava le armi nucleari esistenti, non era valida; era come dire che non valeva la pena prendere provvedimenti per arrestare la corsa agli armamenti se tali provvedimenti non eliminavano completamente gli arsenali nucleari esistenti: una tesi inaccettabile. L'URSS rimase irremovibile nonostante il fatto che la nuova proposta americana ricevesse ampio sostegno da parte dei paesi non allineati.

Dopo l'inizio dei negoziati SALT nel 1969, la questione delle trattative unilaterali per un taglio (anzi, per la maggior parte delle altre misure relative alla limitazione degli armamenti nucleari) fu trattata come se avesse perso il proprio carattere di urgenza e rimase in sospeso. Soltanto una decina d'anni dopo, al tempo cioè della Sessione speciale delle Nazioni Unite per il disarmo, l'idea fu ripresa.

La Sessione speciale delle Nazioni Unite per il disarmo, tenutasi a New York nel maggio e nel giugno del 1978, fu la più grande conferenza del suo genere che fosse mai stata convocata. Vi presero parte 20 capi di governo, quattro vicepresidenti o vice primi ministri e 49 ministri degli esteri. Dopo sei settimane di discussioni e di intense trattative, la sessione raggiunse l'unanimità su un lungo documento finale

Il plutonio 239, un altro isotopo facilmente fissionabile usato nella fabbricazione di armi nucleari, viene generato in un tipico reattore a uranio come prodotto finale di una sequenza di reazioni nucleari in cui un nucleo di uranio 238 del combustibile iniziale cattura un neutrone, diventando uranio 239, che poi decade radioattivamente in due fasi per produrre plutonio 239. A sua volta, un nucleo di plutonio 239 può andar incontro a fissione o può assorbire il neutrone, e in tal caso diventa plutonio 240. (Analogamente avviene per il plutonio 241.) La stabilizzazione della curva della produzione di plutonio 239 significa che alla fine della vita effettiva del carico di combustibile di un tipico reattore ad acqua in pressione questo isotopo si consuma quasi alla stessa velocità alla quale viene prodotto. Il plutonio idoneo alla fabbricazione di armi nucleari, definito in maniera approssimativa come plutonio a elevata concentrazione di plutonio 239, si ottiene in speciali reattori togliendo gli elementi combustibili dal nocciolo del reattore dopo un tempo di combustione relativamente breve. Tuttavia, il plutonio adatto per i reattori (ottenuto dopo una combustione, poniamo, di tre anni) potrebbe essere usato anche per fare esplosivi nucleari; di qui la necessità di inserire apposite clausole in qualsiasi accordo internazionale che ne proibisca la produzione per la fabbricazione di armi nucleari.

contenente una dichiarazione di principi, un programma di azione per il disarmo e la presentazione di un migliore meccanismo procedurale per i negoziati.

Parecchi paesi partecipanti sostennero il taglio come promettente misura di limitazione degli armamenti nucleari. Uno dei fautori più vigorosi fu il primo ministro canadese, Trudeau, il quale delineò una strategia in quattro punti intesa a «soffocare» la corsa agli armamenti nucleari, strategia che prevedeva il bando totale degli esperimenti nucleari, il bando delle prove di volo di tutti i nuovi sistemi di vettori strategici, la messa al bando della produzione di materiale fissile a fini militari e la riduzione delle spese militari per i nuovi sistemi di armi strategiche. A proposito del taglio, egli disse: «L'effetto di questo provvedimento sarebbe quello di porre un limite ben preciso alla disponibilità di materiale idoneo alla fabbricazione di armi nucleari... Esso avrebbe il grande vantaggio di mettere gli stati detentori e quelli non detentori di armi nucleari su un piano di parità molto maggiore rispetto alla situazione prevista fin qui dalle norme del Trattato di non proliferazione.» Nel testo concordato alla fine, il documento finale della sessione speciale chiedeva la «cessazione sia della produzione di armi nucleari di ogni tipo e dei loro vettori, sia della produzione di materiale fissile a fini militari».

Alla sessione regolare dell'Assemblea generale, nell'autunno del 1978, il Canada diede seguito all'iniziativa del suo primo ministro proponendo una mozione sul taglio come misura importante sia per limitare la corsa agli armamenti nucleari, sia per impedire l'ulteriore proliferazione delle armi nucleari. La mozione chiedeva che, in una fase appropriata dei suoi lavori, il Comitato per il disarmo prendesse urgentemente in esame «la cessazione e la proibizione, adeguatamente verificate, della produzione di materiale fissile a fini militari e di altri ordigni esplosivi nucleari». Nella sua enunciazione, il proposto taglio non solo comportava tanto la cessazione della produzione corrente quanto la proibizione di ogni produzione nuova, ma veniva esteso fino a mettere al bando tutti gli esplosivi nucleari «pacifici». L'URSS e gli altri membri del blocco comunista (a eccezione della Romania) si opposero alla mozione poiché essa non andava abbastanza lontano, in quanto non chiedeva anche l'arresto della produzione di armi nucleari.

La massa critica di materiale fissile necessario per produrre un esplosivo nucleare è la quantità minima in cui una reazione a catena può essere sostenuta solo da neutroni veloci. La massa critica dipende da molti fattori, fra cui la purezza fisica del materiale in questione. Per un dato materiale, la massa critica è più piccola quando il materiale fissile è nella sua forma isotopica pura e si trova in mezzo a un'opportuna sostanza capace di riflettere i neutroni, come l'uranio naturale. (Una massa critica di 4,4 chilogrammi di plutonio 239 al 100 per cento, riportata qui in quelle condizioni, sarebbe grande quanto un'arancia.) La variazione della massa critica a seconda della concentrazione di materiale fissile in essa contenuto è meno accentuata per il plutonio che per l'uranio.

Gli USA e il Regno Unito, pur non essendo più come un tempo strenui sostenitori del taglio, votarono a favore della mozione, la quale fu approvata dalla stragrande maggioranza delle nazioni presenti e votanti. Alla sessione del 1979 dell'Assemblea generale, il Canada presentò ancora una mozione che chiedeva al Comitato per il disarmo di continuare a prendere in considerazione il taglio e la mozione fu approvata con una maggioranza di voti ancora più schiacciante di quella del 1978. Tutti questi voti a favore hanno incoraggiato il Canada a continuare a insistere per il taglio, che ancora una volta è in primo piano nell'agenda internazionale fra le numerose questioni relative al disarmo.

Non sono chiare le ragioni per le quali attualmente gli Stati Uniti dovrebbero dimostrare meno entusiasmo per un provvedimento favorevole alla limitazione degli armamenti che essi stessi avevano avviato e sostenuto appassionatamente per più di un ventennio. È possibile ovviamente che, per le migliaia di nuove testate che gli USA potranno fabbricare in base al trattato SALT 2, taluni potranno ritenere inadeguate anche le scorte massicce di materiale fissile idoneo alla fabbricazione di armi nucleari che gli americani hanno accumulato nel corso degli anni. Se però si tiene presente che la riduzione unilaterale della produzione di materiale fissile proposta dagli Stati Uniti nel 1964 equivaleva a un calo della produzione pari al 20 per cento per il plutonio e al 40 per cento per l'uranio molto arricchito e che gli USA si erano offerti inoltre di convertire 60 000 chilogrammi di tale materiale (abbastanza per fabbricare migliaia di testate nucleari) da usi militari a usi pacifici, è difficile credere che adesso gli Stati Uniti possano temere una carenza di materiale esplosivo nucleare. Sembra ragionevole cercare altrove il vero movente.

S'intende che qualsiasi proposta di così ampia portata come la messa al bando della produzione di materiale fissile per armi nucleari è destinata inevitabilmente a suscitare un'opposizione fondata su ingiustizie strategiche percepite o immaginate. Il vecchio spettro dell'URSS che trae profitto dalla situazione farebbe indubbiamente la sua ricomparsa. In realtà sono proprio i sovietici che avrebbero motivo di preoccuparsi per la possibilità che gli USA traggano qualche vantaggio da un eventuale taglio. Al momento della firma del Trattato SALT 2 gli Stati Uniti avevano circa 10 000 testate e bombe strategiche, mentre l'URSS, stando alle cronache, ne aveva circa 5000. In base al trattato, ognuna delle due potenze avrebbe diritto legalmente ad accumulare, di qui al 1985, fino a un massimo di circa 17 000 testate e bombe strategiche, anche se nessuna delle due parti intende apparentemente arrivare a tanto. I piani attuali prevedono che l'arsenale americano aumenti fino a circa 13 000 testate e quello sovietico fino a circa 10 000. (Stando alle cronache, gli USA dispongono anche di circa 22 000 armi nucleari tattiche e l'URSS di un numero leggermente inferiore.)

Né in un caso né nell'altro comunque ci sarebbe veramente ragione di preoccuparsi che da un taglio una delle due parti tragga qualche vantaggio sull'altra. Il Trattato SALT 2 ha fissato un uguale «tetto» complessivo al numero dei vettori strategici e delle testate. Se si vuole che sia accettabile, anche qualsiasi accordo su un taglio dovrebbe fissare un tetto uguale per entrambe le potenze. Se c'è qualche disparità fra le attuali scorte di materiale fissile idoneo alla fabbricazione di armi nucleari, essa potrebbe essere regolata con conversioni e trasferimenti asimmetrici del materiale stesso da scopi militari a scopi pacifici. Pertanto, mentre negli anni cinquanta e sessanta un taglio, unito a un trasferimento a scopi pacifici di materiale militare, non avrebbe contribuito forse a stabilizzare l'equilibrio nucleare fra le due superpotenze, nelle circostanze attuali lo farebbe. Se inoltre la messa al bando iniziale della produzione e i trasferimenti procedessero secondo l'accordo, sarebbe possibile aumentare progressivamente i futuri trasferimenti, e questo potrebbe essere un altro modo per realizzare vere e proprie riduzioni delle armi strategiche nucleari. In questo modo il taglio potrebbe diventare di per sé un passo importante sulla via del conseguimento del disarmo.

È indubbio che si porrebbero degli interrogativi sull'adeguatezza della verifica in una simile messa al bando della produzione. Dati però i notevoli progressi compiuti negli ultimi vent'anni sia nel campo della sorveglianza per mezzo di satelliti artificiali, sia in quello di altri «mezzi tecnici nazionali», la verifica non presenta più gli stessi problemi di una volta dal punto di vista dell'efficacia e dell'ingerenza. Questo fu indubbiamente uno dei fattori che contribuirono alla decisione del presidente americano Nixon di abbandonare nel 1969 l'ispezione dell'avversario e di affidarsi invece ai meno onerosi sistemi internazionali di salvaguardie della IAEA per le attrezzature e gli impianti dichiarati che dovevano essere chiusi.

I moderni mezzi di verifica assicurerebbero che i grandi impianti e le grandi attrezzature occorrenti per produrre quantità significative di uranio fortemente arricchito e di plutonio non possano sfuggire all'identificazione. Anche la diversione segreta di combustibile esaurito dai reattori di potenza alla produzione di plutonio, o lo sviluppo di nuovi processi per arricchire l'uranio, che forse potrebbero permettere la produzione segreta, in piccole installazioni clandestine, di materiale fissile idoneo alla fabbricazione di armi nucleari, non creerebbero gravi instabilità nell'equilibrio nucleare fra le superpotenze. Le quantità che potrebbero essere prodotte segretamente in questo modo sarebbero talmente piccole da non apportare nessuna differenza significativa, ove si consideri l'ordine di grandezza delle scorte americane e sovietiche già esistenti.

Particolare attenzione dovrebbe essere rivolta ai problemi di verifica che sorgerebbero a proposito della produzione di plutonio. Sebbene il plutonio puro 239 sia il più efficiente materiale fissile idoneo alla fabbricazione di armi nucleari, esso può essere fatto esplodere anche quando è in miscela con quantità più piccole di altri isotopi di plutonio fissili e non fissili. Inoltre è estremamente difficile «denaturare» deliberatamente il plutonio 239. Anche quando è associato a ossidi e a isotopi di altri elementi, il plutonio può essere estratto dal combustibile esaurito di un reattore con varie tecniche di separazione chimica. Va da sé che il problema della verifica verrebbe ulteriormente complicato se il mondo dovesse muoversi in direzione di una «economia del plutonio» attraverso la diffusa adozione del riciclaggio del plutonio e dei reattori autofertilizzanti.

Due impostazioni concettuali diverse sono state proposte per affrontare questo problema. Numerosi osservatori hanno suggerito di interrompere ogni ritrattamento del plutonio, una soluzione che di per sé crea molti problemi. L'altra soluzione proposta è quella di porre tutto il plutonio sotto il sistema di salvaguardie della IAEA e di affidare in custodia alla IAEA stessa tutte le scorte di plutonio. Questo richiederebbe il rafforzamento delle salvaguardie della IAEA, che attualmente si basano su sistemi di computo, di contenimento e di sorveglianza e su periodiche ispezioni in loco, e la creazione di speciali attrezzature per l'accumulazione delle scorte. Qualunque sia il metodo adottato, bisognerà negoziare attentamente ogni singolo particolare di qualsiasi sistema che richieda la verifica di un'eventuale messa al bando della produzione di plutonio idoneo alla fabbricazione di armi nucleari o il suo controllo in casi speciali (come il suo uso come combustibile in certi reattori nucleari o in sommergibili o altre imbarcazioni a propulsione nucleare), e si renderanno necessarie particolari norme e precauzioni. Lo stesso discorso varrebbe per l'uranio fortemente arricchito prodotto per particolari scopi non esplosivi. A ogni modo, di fronte a casi dubbi, si potrebbe sempre fare pieno uso dei mezzi tecnici nazionali di verifica e di commissioni consultive, sulla base dei precedenti creati nei vari accordi SALT.

Un altro problema si presenterebbe se, come appare probabile, Cina e Francia non accettassero un taglio. Entrambi questi paesi hanno sostenuto ripetutamente che non avrebbero aderito a una messa al bando totale degli esperimenti finché non avessero colmato o ridotto lo scarto esistente fra loro e le superpotenze in fatto di armi nucleari. Ed è fuor di dubbio che essi assumerebbero la stessa posizione nei confronti di un taglio. La Cina e la Francia per altro sono così indietro in fatto di armi nucleari, che per molti anni il consenso o il dissenso di queste due potenze non farebbe molta differenza dal punto di vista delle superpotenze. Nel 1963 USA, URSS e Regno Unito erano pronti a concludere il Trattato per la messa al bando parziale degli esperimenti e il Trattato di non proliferazione (e a negoziare la messa al bando totale degli esperimenti) senza la partecipazione della Cina e della Francia. Allo stesso modo dovrebbero essere pronti e in grado, senza mettere a repentaglio la propria sicurezza o erodere la propria deterrenza, di accordarsi sul taglio senza la partecipazione della Cina e della Francia, con la speranza che alla fine anche questi due paesi, arroccati sulle proprie posizioni negative, aderiscano all'accordo. Per la verità, qualsiasi accordo per un taglio comprenderebbe quasi certamente quella che ormai è diventata una clausola d'uso nei trattati per il controllo degli armamenti, vale a dire la richiesta di una revisione del funzionamento del trattato dopo un certo numero di anni per valutare in quale misura la realtà abbia in effetti corrisposto alle aspettative.

Si potrebbe sostenere anche che un taglio e l'inizio del disarmo nucleare si risolverebbero in un aumento massiccio delle armi convenzionali, che potrebbe condurre a un grande incremento delle spese militari. Innanzi tutto è bene ripetere che, per pericolose e distruttive che siano, le armi costituiscono una minaccia alla civiltà e alla sopravvivenza dell'uomo; ci vorrebbero anni perché qualsiasi guerra convenzionale provocasse le distruzioni che una guerra nucleare potrebbe invece causare nello spazio di qualche ora o perfino di qualche minuto. Questa tesi inoltre tende a ignorare l'atmosfera enormemente diversa che si verrebbe a creare in tutto il mondo di fronte a qualunque segno di un vero contenimento della corsa agli armamenti nucleari. La conseguente riduzione dei timori e delle tensioni, la ripresa della distensione fra USA e URSS e l'accresciuta fiducia internazionale potrebbero approdare a una maggiore cooperazione internazionale e creare forse condizioni tali da rendere possibili misure di ampia portata non solo nel campo del disarmo convenzionale, ma anche in quello del disarmo nucleare. Naturalmente non c'è nessuna garanzia che il risultato sarebbe questo; la storia però ci ha insegnato che il continuo sviluppo delle armi nucleari non porta al disarmo convenzionale, ma alimenta piuttosto la corsa agli armamenti in entrambe le direzioni.

Si potrebbe sostenere infine che un taglio non porrebbe fine alla corsa agli armamenti nucleari, in quanto USA e URSS potrebbero pur sempre proseguire la corsa sul piano tecnologico, con l'obiettivo di accrescere la precisione e l'efficacia dei loro vettori, e ciò costituirebbe, per la deterrenza e per la sicurezza internazionali, una minaccia ancora più grande di quella posta dalla semplice produzione e accumulazione di un numero maggiore di armi. Di per sé il taglio non interromperebbe la ricerca e lo sviluppo o l'ammodernamento dei sistemi nucleari o

il trasferimento di materiale fissile da tipi più vecchi ad altri più nuovi di testate nucleari. Pertanto, a meno che non vengano promulgate misure più dirette per il disarmo nucleare e per controlli sui nuovi sviluppi tecnologici (come l'interdizione degli esperimenti nucleari effettuati nel sottosuolo), l'influsso della messa al bando della produzione incomincerebbe col tempo a perdere il proprio effetto.

Queste argomentazioni hanno qualche merito. Ovviamente esse andrebbero altrettanto bene a proposito di una messa al bando della produzione di un numero supplementare di armi nucleari. Ciò nonostante, c'è tutta una molteplicità di vantaggi indiretti che si possono avere limitando la quantità di materiale esplosivo nucleare e capovolgendo la tendenza della corsa ad accrescere illimitatamente il numero delle armi nucleari. Come gli USA sostenevano nel 1969, è inaccettabile la tesi secondo la quale provvedimenti intermedi tendenti a porre un freno alla corsa agli armamenti sono privi di qualsiasi utilità se non eliminano completamente tutte le armi nucleari. Per conseguire l'inversione nella direzione di marcia di una corsa è necessario innanzi tutto fermarla. Inoltre l'atto stesso di porre fine alla produzione di materiale esplosivo necessario per produrre altre armi potrebbe creare un clima più favorevole alla cessazione della produzione di altre armi nucleari.

La riuscita verifica di un taglio avrebbe grande importanza politica e psicologica. Potrebbe incidere profondamente sulle idee convenzionali, secondo le quali non si può fare nulla per fermare la corsa agli armamenti. Avrebbe qualche effetto nel ridimensionare l'importanza delle armi nucleari nelle relazioni internazionali e potrebbe portare a una significativa riduzione delle tensioni attualmente esistenti fra USA e URSS. Essa potrebbe dare anche una dimostrazione delle possibilità di verifica sul piano internazionale che sarebbe utile per altre misure relative al controllo degli armamenti nucleari e al disarmo.

Attualmente 110 paesi aderiscono al Trattato di non proliferazione, mentre più di 40 ne sono fuori. Questi ultimi per la maggior parte non hanno né la capacità né il desiderio di «diventare potenze nucleari», ma alcuni di essi hanno già tale capacità o l'avranno entro breve tempo. Fra questi vi sono non soltanto l'India, che ha già fatto esplodere un ordigno nucleare, ma anche il Pakistan, Israele, l'Egitto, il Sud Africa, la Spagna, l'Argentina e il Brasile. Fra i paesi aderenti al Trattato di non proliferazione, circa tre dozzine sono potenziali detentori di armi nucleari. Secondo le clausole del trattato, essi possono ritirarsi con un preavviso di tre mesi se hanno l'impressione che i loro interessi vitali siano messi in pericolo. La possibilità o meno che qualcuno di essi decida di ritirarsi e si appresti a diventare una potenza nucleare dipenderà in parte dall'eventualità che qualche (e quale) altro stato non detentore di armi nucleari decida di entrarne in possesso. È quindi d'importanza fondamentale per tutti gli stati interessati al rafforzamento del regime di non proliferazione fare tutto quanto è possibile per ridurre gli incentivi a diventare potenze nucleari.

Vi sono molte cose che gli stati detentori di armi nucleari possono fare a questo proposito. Una consiste nel porre termine alla proliferazione verticale, iniziativa che però richiederà qualche tempo. Un passo più immediato sarebbe quello di eliminare, o ridurre al massimo, gli elementi discriminatori del Trattato di non proliferazione. Quando questo accordo veniva negoziato negli anni sessanta, molti dei paesi sprovvisti di armi nucleari guardavano con preoccupazione all'intrinseco squilibrio del trattato, il quale decretava una divisione permanente degli stati del mondo in due categorie: quelli dotati e quelli non dotati di armi nucleari. Quasi tutti i paesi erano disposti ad accettare la situazione come fatto temporaneo della vita internazionale, a condizione che gli stati in possesso di armi nucleari si impegnassero a fermare e a invertire la tendenza della corsa agli armamenti nucleari (in conformità all'articolo VI del trattato) e ad aiutare quegli stati, che di armi nucleari erano sprovvisti, nello sviluppo degli usi pacifici dell'energia nucleare (in conformità all'articolo IV). Il problema della mancanza di uguali diritti e doveri fra gli stati detentori e quelli non detentori di armi nucleari fu reso più difficile dalle clausole del trattato che imponevano soltanto agli stati che non possedevano armi nucleari di sottoporre in blocco i loro programmi nucleari al sistema di salvaguardie della IAEA.

Ognuno dei paesi potenzialmente detentori di armi nucleari, indicati qui in varie sfumature di colore, potrebbe avere entro il 1990 la capacità tecnica di far esplodere un ordigno nucleare. I paesi in colore più intenso potrebbero acquisire tale capacità entro un anno o due, quelli in colore medio entro due-sei anni e quelli in colore più chiaro entro sei-dieci anni. Le nazioni che hanno già compiuto esperimenti nucleari sono rappresentate in grigio. Gli asterischi indicano le nazioni che non hanno aderito al Trattato di non proliferazione. Sebbene né Israele né Sud Africa abbiano fatto esplodere un ordigno nucleare, risulta che tanto l'uno quanto l'altro hanno già la capacità di farlo. La Libia è fra gli stati potenzialmente detentori di armi nucleari in virtù dei suoi legami col Pakistan. Questa è una versione aggiornata di una cartina pubblicata dall'autore dell'articolo sul numero 84, agosto 1975, della rivista.

Tale e tanto fu il risentimento dei paesi in questione di fronte a quella che per molti di essi era una situazione iniqua e discriminatoria, che USA e Regno Unito (ma non l'URSS) stabilirono volontariamente di accettare le salvaguardie della IAEA sui loro programmi e sulle loro attrezzature nucleari non militari. Questa concessione andava, sì, nella direzione del conseguimento di un maggiore equilibrio fra le due categorie di stati, ma non si spingeva abbastanza avanti da soddisfare paesi come l'India, il Brasile e l'Argentina, agli occhi dei quali il trattato non aveva perso alcuna caratteristica del suo iniziale carattere discriminatorio.

Perfino gli stati non detentori di armi nucleari che hanno aderito al Trattato di non proliferazione sono insoddisfatti di quello che essi considerano il mancato adempimento, da parte dei paesi che le armi nucleari le possiedono, degli impegni che essi si erano assunti in base alle clausole del trattato. Nel 1975, alla prima Conferenza di revisione del Trattato di non proliferazione gli stati non allineati e non detentori di armi nucleari (il cosiddetto Gruppo dei 77) insistettero nel pretendere che i paesi in possesso di armi nucleari tenessero fede ai loro impegni e dessero inoltre garanzie di sicurezza agli stati non detentori, impegnandosi a non fare uso o a non minacciare di fare uso delle armi nucleari contro di loro. La dichiarazione di compromesso che alla fine fu approvata all'unanimità lasciò insoddisfatti quasi tutti gli stati non detentori di armi nucleari.

Nei cinque anni che sono trascorsi dalla prima Conferenza di revisione del trattato di non proliferazione non si sono fatti molti progressi sulla via indicata dalle richieste dei paesi non dotati di armi nucleari. Anche se gli stati membri del Nuclear Suppliers Group, che si incontrarono a Londra dopo il 1975, riuscirono, di propria iniziativa e nel proprio interesse, a mettersi d'accordo per rafforzare le restrizioni e le salvaguardie sull'esportazione di materiale e di apparecchiature nucleari, i paesi sprovvisti di armi nucleari mostrarono la generale tendenza a considerare questo comportamento come uno sforzo di parte inteso ad accentuare gli aspetti discriminatori del regime di non proliferazione.

L'International Nuclear Fuel Cycle Evaluation (INFCE) ha appena portato a termine un'inchiesta particolareggiata, svolta nell'arco di due anni con la partecipazione sia dei paesi importatori sia di quelli fornitori, sui molti problemi dell'energia nucleare e sul loro rapporto con la proliferazione delle armi nucleari. La relazione dell'INFCE, messa in circolazione in febbraio, contiene moltissime e preziose analisi e informazioni. Si tratta però di uno studio non politico, ma rigorosamente tecnico. Di conseguenza non offre nessuna soluzione ai principali problemi dell'energia nucleare e della proliferazione delle armi nucleari, che sono in gran parte di natura politica. Solo una combinazione di misure tecniche e politiche può dare qualche speranza di riuscire ad arrestare un'ulteriore proliferazione delle armi nucleari.

Delle misure politiche necessarie, molte sono legate all'arresto e all'inversione di tendenza della proliferazione verticale delle armi nucleari, intese come modo per facilitare gli sforzi tendenti a impedirne la proliferazione orizzontale. In questo contesto i provvedimenti immediati più importanti che si potrebbero prendere sarebbero l'interdizione totale degli esperimenti e il taglio. L'influsso di una messa al bando totale degli esperimenti è ovvio: essa metterebbe gli stati detentori e quelli non detentori di armi nucleari su un piano di maggiore parità, se a entrambi fosse fatto divieto di condurre esplosioni nucleari tanto di natura militare quanto di natura pacifica; costituirebbe un controllo efficace sul miglioramento tecnologico delle testate nucleari, e avrebbe, infine, un effetto simbolico e psicologico immediato, dimostrando la buona volontà degli stati detentori di armi nucleari di accettare un limite alla corsa qualitativa agli armamenti nucleari.

L'influsso del taglio, che sarebbe analogo a quello dell'interdizione totale degli esperimenti, richiede qualche precisazione. È ovvio che ci sarebbe sempre la possibilità che i paesi in possesso di armi nucleari producano altre testate nucleari attingendo dalle proprie scorte di materiale fissile, o trasferiscano materiale esplosivo dalle testate più grandi a quelle più piccole o da quelle più vecchie a quelle più nuove. Il taglio costituirebbe però pur sempre un passo importante verso il controllo della proliferazione verticale; esso rallenterebbe la produzione di armi nucleari e potrebbe essere una mossa molto utile per porre fine una volta per tutte alla produzione dei vari tipi di armi nucleari. In tal modo contribuirebbe a ridurre gli aspetti discriminatori del Trattato di non proliferazione.

Oltre tutto, una volta arrestata la produzione di materiale fissile a fini militari, la futura produzione di materiale non potrebbe che essere destinata a scopi civili pacifici. Come fu messo in evidenza dagli Stati Uniti nel 1969, ciò renderebbe possibile far accettare ai paesi detentori di armi nucleari le stesse salvaguardie

IAEA che venivano imposte a tutti gli altri paesi in base al Trattato di non proliferazione, in quanto ci sarebbero pochi segreti militari da proteggere. Così, d'un tratto, potrebbe essere rimosso uno dei più irritanti elementi di discriminazione fra gli stati detentori e quelli non detentori di armi nucleari. Particolari precauzioni si renderanno necessarie per evitare la divulgazione di certi segreti dell'arricchimento dell'uranio, ma non è detto che tali precauzioni debbano essere molto diverse da quelle che vengono adottate attualmente nell'ambito delle salvaguardie IAEA per impedire la rivelazione di segreti tecnologici e commerciali relativi ad attrezzature e a reattori nucleari di uso pacifico. Se fossero necessarie procedure supplementari di verifica per assicurare che negli stati detentori di armi nucleari gli impianti di arricchimento dell'uranio e di ritrattamento del plutonio non vengano usati per produrre materiale fissile per armi nucleari, queste procedure sarebbero un necessario fattore concomitante delle maggiori capacità nucleari dei paesi detentori di armi nucleari e non ridurrebbero la sostanziale equità del trattato fin tanto che le salvaguardie imposte agli stati detentori di armi nucleari non fossero meno meticolose e meno efficaci di quelle imposte agli stati non detentori.

Facendo un passo concreto sulla strada della limitazione e della riduzione degli armamenti nucleari, offrendo la speranza di ulteriori misure di disarmo nucleare ed eliminando la diversità di trattamento in fatto di salvaguardie internazionali, la proposta messa al bando della produzione potrebbe contribuire a rafforzare il Trattato di non proliferazione e a eliminare, o per lo meno indebolire, alcune delle argomentazioni addotte contro di esso. Se fosse negoziato come trattato multilaterale, un accordo per un «taglio» costituirebbe una diretta barriera legale alla diffusione delle armi nucleari ad altri paesi che aderissero all'accordo stesso. Per essere valido anche nei confronti degli stati non detentori di armi nucleari impegnati nella produzione di materiale fissile, un trattato del genere dovrebbe prevedere la proibizione oltre che la cessazione della produzione di materiale fissile eventualmente utilizzabile per la fabbricazione di armi e di altri ordigni esplosivi nucleari.

Una proibizione di questo genere andrebbe notevolmente al di là delle clausole del Trattato di non proliferazione e dell'imminente interdizione totale degli esperimenti, in quanto imporrebbe in anticipo una barriera legale alla diffusione delle armi e degli ordigni esplosivi nucleari ai paesi che ancora non ne sono in possesso. Se il trattato dovesse prescrivere in maniera equa l'applicazione delle salvaguardie sia agli stati detentori sia a quelli non detentori di armi nucleari, e se dovesse anche contemplare la conversione e il trasferimento a usi civili pacifici delle scorte esistenti di materiale fissile idoneo alla fabbricazione di armi nucleari, soprattutto a favore dei paesi non dotati di armi nucleari, questi ultimi troverebbero in queste disposizioni un incentivo per aderire al trattato.

Vi sono parecchie e svariate vie alternative che si potrebbero seguire nel negoziare una messa al bando della produzione di materiale fissile per armi nucleari. L'ideale sarebbe di poterla negoziare sin dall'inizio nell'ambito di un trattato multilaterale a cui partecipassero tanto gli stati detentori quanto quelli non detentori di armi nucleari. Dal punto di vista pratico però potrebbe forse risultare necessario incominciare con un negoziato bilaterale, fra le due superpotenze, o trilaterale, includendo anche il Regno Unito. Qualsiasi accordo preliminare raggiunto in questo modo potrebbe essere sottoposto poi al Comitato di Ginevra per il disarmo, in modo da dare agli stati non detentori di armi nucleari l'opportunità di discutere a fondo e di riesaminare il trattato, dopo di che l'accordo potrebbe essere sottoposto al vaglio dell'Assemblea generale delle Nazioni Unite. In questo modo tutti gli stati avrebbero la possibilità di partecipare all'elaborazione del trattato e questa procedura potrebbe facilitare la più ampia adesione possibile al trattato stesso.

Le procedure di verifica attinenti agli stati detentori di armi nucleari potrebbero riguardare in un primo momento la chiusura completa di tutti gli impianti su basi reciproche o comparabili, così come era stato proposto dagli Stati Uniti negli anni sessanta, servendosi delle tecniche di verifica disponibili, ivi compresi i vari mezzi tecnici nazionali. Le procedure più complesse necessarie per assicurare la conversione a usi pacifici del materiale fissile idoneo alla fabbricazione di armi nucleari potrebbero essere verificate col sistema di salvaguardie della IAEA, che attualmente, in base al Trattato di non proliferazione, controlla tutte le attrezzature nucleari degli stati non detentori di armi nucleari. Procedure speciali, maggiormente rigorose, dovrebbero essere negoziate per gli stati non detentori di armi nucleari in cooperazione con la IAEA a proposito dei vari settori problematici discussi nelle righe precedenti. Dovrebbero essere elaborate analogamente procedure per la conversione a scopi pacifici delle scorte militari.

Un'altra alternativa aperta agli USA sarebbe quella di assumere unilateralmente l'iniziativa di un taglio. Dato il ritmo esasperatamente lento dei negoziati relativi a qualsiasi accordo internazionale e date le complicazioni che sorgono quando si tratta di ottenere l'approvazione del senato americano, tutto depone a favore di un'azione unilaterale. Un'azione simile, intesa a facilitare un'azione reciproca internazionale o negoziati, presenta anche molti precedenti. Gli Stati Uniti potrebbero benissimo prendere in considerazione l'idea di ripetere la loro iniziativa del 1964, quando annunciarono unilateralmente una riduzione della loro produzione di materiale fissile a fini militari e invitarono l'URSS a fare altrettanto. Il gesto americano portò a riduzioni reciproche annunciate dall'URSS e dal Regno Unito, che furono definite «atti unilaterali contraccambiati» o, come dissero i sovietici, azione per «mutuo esempio».

Con un gesto analogo il presidente americano Nixon annunciò nel 1969 la decisione unilaterale degli Stati Uniti di rinunciare a ogni uso di armi e di agenti batteriologici, di interrompere la produzione di tutti gli agenti biologici a fini militari, di distruggere tutte le scorte esistenti di armi biologiche e di limitare a scopi difensivi tutta la futura ricerca biologica. Gli Stati Uniti speravano che anche altri stati rinunciassero alle armi biologiche ed esercitarono pressioni affinché tali decisioni fossero convertite in obblighi legali internazionali attraverso una convenzione di qualche genere. Questo gesto unilaterale degli Stati Uniti portò al negoziato e al felice esito nel 1972 della Convenzione per la proibizione dello sviluppo, della produzione e dell'accumulazione di scorte di armi batteriologiche (biologiche) e di armi a base di tossine e per la loro distruzione.

Non sembra proprio che un blocco unilaterale della produzione di materiale fissile a usi militari, per lo meno sotto forma di una sospensione, possa mettere gli USA in una posizione di svantaggio rispetto all'URSS, soprattutto se l'iniziativa fosse presa su basi sperimentali temporanee e se i sovietici fossero invitati a compiere un analogo gesto unilaterale e ad avviare negoziati per un trattato su un «taglio».

Non è quindi da escludere che gli Stati Uniti prendano in considerazione la possibilità, con questa azione unilaterale, di promuovere la causa del contenimento della proliferazione delle armi nucleari sia in senso verticale sia in senso orizzontale. Se fosse preparata accuratamente in segreto attraverso i canali diplomatici, probabilmente l'iniziativa potrebbe avere lo stesso esito positivo della riduzione unilaterale della produzione di materiale fissile a fini militari decisa nel 1964 e del blocco unilaterale della produzione di agenti biologici a fini militari annunciato nel 1969. Un'azione nazionale unilaterale di questo genere presenta il notevole pregio della rapidità e della tempestività tanto in fase di avvio quanto in fase di attuazione.

L'anno scorso, nella loro Dichiarazione congiunta di princìpi e di linee di condotta fondamentali per successivi negoziati sulla limitazione degli armamenti strategici, le due superpotenze affermarono di comune accordo la volontà di perseguire «significative e considerevoli riduzioni del numero delle armi strategiche offensive (e) limitazioni qualitative, comprese restrizioni in materia di sviluppo, sperimentazione e schieramento di nuovi tipi di armi strategiche offensive e di ammodernamento delle armi strategiche offensive già esistenti». Ovviamente è nel loro reciproco interesse agire in questo modo, non solamente inteso come mezzo per frenare la corsa agli armamenti, ma anche come possibile alternativa per limitare la diffusione delle armi nucleari ad altri paesi.

Quando il taglio fu proposto per la prima volta dal presidente americano Eisenhower nel 1953, la situazione mondiale era molto diversa da quella attuale. L'URSS era molto indietro rispetto agli Stati Uniti in tutti i settori dell'armamento nucleare e il pericolo della proliferazione delle armi nucleari era ancora piuttosto remoto. Oggi le due superpotenze sono più o meno su un piano di parità e la minaccia dell'ulteriore diffusione delle armi nucleari ad altri paesi, con tutti gli imprevedibili problemi che essa comporterebbe per le superpotenze e per la relativa stabilità della situazione internazionale, è un pericolo molto reale.

Non c'è nessuna garanzia che, se le superpotenze riuscissero a mettersi d'accordo sulla messa al bando della produzione di materiale fissile idoneo alla fabbricazione di bombe o di altri ordigni esplosivi nucleari, la cosa basterebbe da sola a impedire la proliferazione delle armi nucleari ad altri paesi. Ciò nonostante, un trattato per un taglio, soprattutto se fosse seguito da vicino da un'interdizione totale degli esperimenti, potrebbe dare una prova concreta del fatto che le superpotenze sono veramente decise ad arrestare la corsa agli armamenti nucleari e a invertirne la tendenza. Esso forse potrebbe oltre tutto creare un chiaro e sufficiente effetto di retroazione che faciliterebbe altre misure intese a frenare tanto la proliferazione verticale quanto quella orizzontale.

A parte le considerazioni di maggiore rilievo delineate in questo articolo, c'è ancora un'altra ragione per procedere a un taglio. Nonostante l'opposizione alle centrali nucleari che si è andata creando negli ultimi anni, sembra abbastanza scontato che in futuro il mondo dovrà per forza di cose affidarsi in misura sempre maggiore all'energia nucleare come fonte importante e probabilmente crescente di energia. I timori di alcuni anni or sono a proposito di una possibile carenza di uranio per i reattori nucleari, che per il momento sembrano accantonati, potrebbero benissimo ridestarsi.

L'uranio molto arricchito di cui si sono accumulate ingenti scorte per la fabbricazione di armi potrebbe costituire una fonte di uranio leggermente arricchito per i reattori di potenza. 1000 chilogrammi di uranio idoneo alla fabbricazione di armi potrebbero essere convertiti in 30 000 chilogrammi di combustibile nucleare per i reattori. Questa fonte di energia contribuirebbe di per sé a differire l'avvento dei reattori autofertilizzanti d'uso commerciale e a rinviare anche l'economia del plutonio. Se, inoltre, tutto il materiale fissile prodotto in futuro dovesse essere usato esclusivamente a scopi pacifici, non dovrebbe esserci più tanta paura che, nel caso in cui il mondo si muovesse effettivamente in direzione dei reattori autofertilizzanti e del ritrattamento del plutonio, la cosa porterebbe necessariamente alla proliferazione delle armi nucleari.

C'è ovviamente la questione sempre attuale della politica del «taglio». Le difficoltà di giungere a un accordo sul Trattato SALT 2 e l'incombente minaccia della mancata ratifica da parte del senato americano sono così grandi da sollevare dubbi perfino tra i fautori più ottimisti del controllo degli armamenti. Lo stato d'animo del pubblico e del Congresso degli Stati Uniti attualmente rasenta l'ostilità nei confronti di misure favorevoli al controllo degli armamenti. Questa avversione però rende tanto più necessaria la ricerca di nuove idee e di nuove opportunità che tengano in vita le trattative. Nelle circostanze più pessimistiche, ove si eccettui la guerra, è quasi inconcepibile che le due superpotenze abbandonino i propri sforzi per arrestare la corsa agli armamenti nucleari e contenere la proliferazione delle armi nucleari. Anche durante la «guerra fredda» gli USA ebbero iniziative ingegnose, come il piano Baruch del 1946 e la proposta «Atomi per la pace» del 1953, rispetto alle quali qualunque iniziativa a favore del taglio sarebbe estremamente più semplice e di gran lunga meno radicale.

Il fatto che in passato l'URSS si sia opposta alla sospensione della produzione di materiale fissile a fini militari senza una corrispondente sospensione della produzione di armi nucleari non dovrebbe essere considerato un ostacolo insormontabile. Nel caso di quasi tutti gli accordi multilaterali per la limitazione degli armamenti, l'URSS sulle prime ha sempre assunto una posizione radicale e chiesto proibizioni totali. Quando però era ormai chiaro che la stragrande maggioranza degli stati era contraria a un'impostazione del tipo «o tutto o nulla» e preferiva quella del tipo «meglio poco che niente», l'URSS ha accettato in quasi tutti i casi una soluzione parziale, spesso sulla falsariga delle proposte americane. Lo stesso modello di comportamento si è ripetuto in molte occasioni nel corso dei negoziati SALT. Perciò, in base non solo alla logica, ma anche all'esperienza maturata col tempo, vi sono tutte le ragioni per sperare che l'Unione Sovietica accetti anche il taglio come primo passo verso ulteriori misure di limitazione degli armamenti nucleari.

La recente iniziativa del Canada ha già avuto l'appoggio di quasi tutti i paesi occidentali e di quelli non allineati. Se i leader dell'URSS pongono mente ancora una volta al taglio, il problema potrebbe essere in via di soluzione. A parte l'imminente traguardo di un Trattato per l'interdizione totale degli esperimenti, è difficile immaginare una iniziativa nuova più promettente di un trattato che metta al bando la produzione di materiale fissile per armi nucleari. Uno sforzo in questa direzione potrebbe fare molto, allo stesso tempo, per restituire un po' di credibilità, attualmente sempre più in ribasso, all'impegno americano in merito al controllo e alla non proliferazione degli armamenti nucleari.

Astronomia di righe gamma

Strumenti a bordo di palloni e satelliti consentono di rilevare le righe spettrali di radiazioni elettromagnetiche emesse da processi astrofisici, finemente accordate a energie anche più grandi di quelle dei raggi X

di Marvin Leventhal e Crawford J. MacCallum

La struttura dell'universo è stata dedotta quasi esclusivamente dai fotoni (cioè i quanti di energia elettromagnetica) che giungono in vicinanza della Terra. Fino a circa mezzo secolo fa gli astronomi potevano rivelare solo fotoni con energie comprese fra 1,5 e 3,5 elettronvolt circa: i fotoni della luce visibile. Successivamente essi incominciarono a estendere l'intervallo di energia dei fotoni rivelabili verso il basso, nelle regioni dell'infrarosso e delle onde radio e verso l'alto, nella regione dell'ultravioletto vicino. Con l'avvento dei razzi, dei palloni da grande altezza e dei satelliti, essi riuscirono a estendere ulteriormente tale intervallo verso l'alto, nella regione dei fotoni che non possono attraversare l'atmosfera terrestre: i fotoni dell'ultravioletto lontano, i raggi X e i raggi gamma.

I raggi gamma costituiscono la forma di radiazione elettromagnetica di energia più elevata; l'energia di ciascuno dei loro fotoni si misura in milioni di elettronvolt (MeV), e in linea di principio non ha un limite superiore. I primi raggi gamma dallo spazio furono scoperti circa una ventina di anni fa. I primi rivelatori registravano semplicemente l'arrivo dei fotoni, ma non erano in grado di analizzarne l'energia, come avviene per i fotoni della luce visibile, che vengono analizzati in righe spettrali mediante spettroscopi. Ma oggi sono disponibili strumenti che possono rivelare righe spettrali di raggi gamma, e grazie a essi si incominciano a ricavare informazioni su oggetti astronomici e processi di alta energia che dominano l'attenzione degli astronomi, come per esempio le supernove, le stelle di neutroni e i fenomeni che si svolgono nel centro delle galassie.

Mentre le righe dello spettro visibile derivano da transizioni fra i livelli energetici degli elettroni negli atomi, le righe dello spettro gamma derivano da transizioni fra i livelli energetici dei nuclei atomici. Consideriamo per esempio il nucleo del neo 22, che è formato da 10 protoni e 12 neutroni. Il nucleo si trova di solito nel suo stato fondamentale, cioè nel suo livello energetico più basso, ma può anche risiedere in uno dei suoi livelli di energia eccitati. Se si trova nel livello eccitato più basso, decade nel livello fondamentale in un tempo dell'ordine di 10^{-12} secondi. La transizione è accompagnata dall'emissione di un fotone gamma con una energia di esattamente 1,2746 MeV, che corrisponde alla differenza di energia fra i due stati. Il raggio gamma costituisce dunque la «firma» di una specifica transizione nucleare.

La transizione può avvenire come conseguenza di molti processi distinti. In uno di questi casi un nucleo di sodio 22, che è radioattivo, decade in uno stato eccitato del neo 22 con periodo di dimezzamento di 2,6 anni. Si pensa che questo decadimento sia un processo comune nell'evoluzione di una supernova. In un altro caso una particella subatomica di alta energia urta contro un nucleo di neo 22 nel suo stato fondamentale e lo eccita a un livello energetico più elevato. In un terzo tipo di processo la cattura di un neutrone da parte di un nucleo di neo 21 genera un nucleo eccitato di neo 22. In ciascuno di questi casi il decadimento - in realtà quasi immediato - del nucleo eccitato al suo stato fondamentale è accompagnato dall'emissione di un fotone gamma di energia ben determinata.

Righe gamma possono derivare anche da processi che non coinvolgono nuclei atomici. Uno di questi processi è l'urto fra un elettrone e un positone (o antielettrone). Le due particelle hanno proprietà identiche, fatta eccezione per la carica elettrica, che è negativa per l'elettrone e positiva per il positone. Quando un elettrone e un positone si incontrano, si annichilano l'un l'altro. La loro massa viene convertita in energia elettromagnetica, che si irradia nello spazio dal punto di annichilazione, generalmente sotto forma di due fotoni gamma, ciascuno con una energia di 0,511 MeV. L'annichilazione elettrone-positone sembra essere un processo molto comune nelle stelle di neutroni e nel centro della Galassia.

Un altro processo non nucleare che dà origine a righe gamma è la radiazione di ciclotrone degli elettroni. Gli elettroni che si muovono entro un campo magnetico sono soggetti a una accelerazione centripeta che provoca l'irraggiamento di fotoni a una energia caratteristica. Questa energia è proporzionale all'intensità del campo, e pertanto, se il campo è sufficientemente intenso e uniforme (come potrebbe essere presso i poli magnetici di una stella di neutroni) gli elettroni irraggiano fotoni gamma.

Le prime righe gamma provenienti dallo spazio e identificate con certezza erano attribuibili a diversi fra questi meccanismi. Esse provenivano da una gigantesca facola solare prodottasi nell'agosto del 1972, e furono scoperti da Edward L. Chupp e dai suoi colleghi dell'Università del New Hampshire, che avevano collocato un piccolo rivelatore di raggi gamma a bordo del satellite *OSO-7* (il settimo Osservatorio Solare Orbitante della National Aeronautics and Space Administration). Fra le righe era presente anche quella da 0,511 MeV, derivante dalla annichilazione, nell'atmosfera solare, di positoni prodotti nella facola. Un'altra riga, a 2,2 MeV, derivava dalla cattura di neutroni da parte di protoni, con l'emissione di un raggio gamma da parte di ogni nucleo di deuterio così formato. Infine, deboli righe a 4,4 e 6,1 MeV derivavano dal decadimento allo stato fondamentale di nuclei di carbonio 12 e di ossigeno 16 che erano stati portati a livelli energetici eccitati da urti, nell'atmosfera solare, contro protoni accelerati nella facola.

Negli otto anni trascorsi da quelle prime osservazioni diversi gruppi di ricerca, incluso il nostro, operante presso i Bell Laboratories e i Sandia Laboratories, hanno fatto volare telescopi per raggi gamma, nella maggior parte dei casi con palloni, ma talvolta su satelliti, nel tentativo di portare gli strumenti al di sopra di quasi tutta l'atmosfera terrestre e di rivelare righe gamma di origine astrofisica. Questo campo di ricerca è ancora nella sua infanzia; qui passeremo in rassegna alcune delle prospettive del settore, descriveremo un telescopio per raggi gamma e riferiremo alcuni dei primi risultati ottenuti.

La più importante, forse, fra le prospettive, è quella di raccogliere osservazioni che si riferiscono alla sintesi cosmica degli elementi chimici. Si fa infatti l'ipotesi che il big bang iniziale abbia creato soprattutto idrogeno ed elio, che vengono poi trasformati negli elementi più pesanti per fusione nucleare nell'interno delle stelle. Ma per gli elementi più pesanti del ferro 56 la fusione non è esoenergetica. Per spiegare la genesi di questi elementi bisogna perciò invocare un qualche meccanismo diverso dalla fusione nell'interno delle stelle: più specificamente, si fa l'ipotesi che gli elementi intermedi e quelli più pesanti siano prodotti principalmente nelle esplosioni delle stelle.

I dettagli sono i seguenti. Una stella è una sfera di gas nella quale il calore generato dalla fusione nucleare contrasta la tendenza del gas a cadere verso il centro. Ma alla fine, la stella esaurisce la sua riserva di combustibile nucleare. Una ipotesi bene accetta sostiene che se la massa della stella è dell'ordine di quella del Sole, la stella si contrae gradualmente fino ad avere approssimativamente il raggio della Terra. Ora è una nana bianca: ha esaurito il combustibile nucleare e sta raffreddandosi. Se invece la massa è più grande, la stella non si contrae gradualmente, ma implode catastroficamente e quindi esplode. Si calcola che una volta, in un periodo che va da 10 a 300 anni, una stella fra le centinaia di miliardi che costituiscono una tipica galassia raggiunge questo punto finale della evoluzione stellare, e la luce di questa esplosione può eclissare l'intera galassia per un periodo di diversi giorni o addirittura settimane. È questo l'evento chiamato supernova. Esso lascia dietro di sé un insieme di frammenti

Il lancio, all'alba del 9 ottobre 1975, del pallone che trasportò il telescopio a righe gamma degli autori in un volo al di sopra di Alamogordo, New Mexico. Il pallone è attaccato a un paracadute, agganciato all'argano a sinistra, a cui è sospeso il carico utile. La base del carico comprende un insieme di ammortizzatori a schiacciamento. Il cono al di sotto degli ammortizzatori è zavorra. La scatola in cima al carico contiene circuiti elettronici e di fronte c'è un bidone di azoto liquido usato come refrigerante. Il telescopio è situato fra la scatola e il bidone, per cui non è visibile nella fotografia. Alla fine del volo il paracadute con il suo carico verrà sganciato dal pallone.

in espansione, di cui l'esempio più noto nella nostra galassia è la Nebulosa del Granchio nella costellazione del Toro.

Gli astrofisici teorici W. David Arnett, Donald D. Clayton e Stirling A. Colgate furono fra i primi a costruire modelli matematici delle reazioni nucleari che avvengono in quei pochi esplosivi secondi, in cui i processi di nucleosintesi stellare, che fino ad allora si sono svolti presumibilmente per miliardi di anni, assumono andamenti completamente nuovi. Questi modelli suggeriscono che l'esplosione tende a formare nuclei più pesanti a partire dai più leggeri. In particolare sembra in effetti possibile che le supernove abbiano creato la maggior parte degli elementi intermedi e pesanti, con le abbondanze cosmiche relative che vengono osservate.

Naturalmente sarebbe bene sottoporre queste teorie a una prova, e in linea di principio ciò può essere fatto. Le righe spettrali a energie ottiche e quelle di raggi X rivelano quali tipi di atomi siano presenti nei resti di una esplosione di supernova. Gli atomi che emettono le radiazioni, tuttavia, potrebbero aver fatto parte della stella prima dell'esplosione, e alcuni di essi potrebbero essere stati raccolti dallo spazio interstellare. D'altra parte molti dei nuclei sintetizzati nell'esplosione dovrebbero essere radioattivi, e il decadimento di tali nuclei è accompagnato dall'emissione di raggi gamma con energie caratteristiche (alcune delle radiazioni caratteristiche sono a energie un po' più basse e corrispondono a raggi X di alta energia). La scoperta anche di una sola riga gamma proveniente dalla posizione di una esplosione di supernova costituirebbe perciò un progresso molto importante nella nostra comprensione di come si è formato l'attuale universo.

Un secondo tipo di esplosione stellare è quello noto col nome di nova. Le nove sono molto meno violente delle supernove, ma molto più frequenti; si verificano forse fino a 25 esplosioni del genere all'anno per galassia. Anch'esse possono dare un contributo importante alla nucleosintesi. Nello schema correntemente accettato una esplosione di nova si verifica in un sistema stellare binario in cui una delle stelle è una stella ordinaria e l'altra è una nana bianca, una stella che, come abbiamo visto, ha esaurito il combustibile nucleare senza esplodere. Benché una nana bianca non abbia una sorgente di energia propria, si pensa che quando una stella di questo tipo appartiene a un sistema binario si aggrega lentamente materia, principalmente idrogeno, sottraendola alla stella gemella. Prima o poi la temperatura e la pressione alla base di questo involucro gassoso crescente raggiungono valori esplosivi e l'involucro scoppia.

I calcoli eseguiti per la nucleosintesi sulle esplosioni di nove da gruppi guidati da Clayton e da James W. Truran, Jr., suggeriscono una grande produzione di sodio 22, che decade per la quasi totalità in uno stato eccitato del neo 22, con una vita media di 2,6 anni. Diversi mesi dopo l'esplosione la riga gamma più intensa proveniente dal luogo dell'esplosione dovrebbe perciò essere la riga a 1,2746 MeV, che rappresenta il decadimento dei nuclei di neo, appena formati, al loro stato fondamentale. Alcuni mesi dovrebbero essere sufficienti per gli astronomi per spedire in aria un telescopio per raggi gamma portato da un pallone. Poiché le nove sono frequenti, e poiché è probabile che qualcuna di esse sia vicina al sistema solare, potrebbe ben avvenire che la riga a 1,2746 MeV sia la prima riga gamma rivelata proveniente dall'esplosione di una stella.

Un nucleo atomico emette un raggio gamma quando decade da un livello di energia eccitato allo stato fondamentale, che è il livello di minima energia possibile. In questa figura sono schematizzati i tre possibili modi in cui può avvenire l'emissione da parte di un nucleo di neo 22. Nella cattura neutronica *(riga in alto)* un nucleo di neo 21 si trasforma in un nucleo di neo 22 eccitato mediante l'aggiunta di un neutrone *(n)*. Nel processo di decadimento radioattivo *(seconda e terza riga)*, il sodio 22 diventa neo 22 in uno stato eccitato mediante l'aggiunta di un elettrone (e^-), generalmente per cattura di uno degli elettroni dell'atomo stesso, o mediante l'emissione di un positone (e^+), che è l'antiparticella dell'elettrone. Il secondo processo è il più frequente dei due, ed entrambi portano alla trasformazione di un protone in un neutrone. Nell'eccitazione per urto *(riga inferiore)* l'energia trasmessa al neo 22 da una particella veloce (in questo caso un protone *p*) lo fa passare in uno stato eccitato. In tutti i casi il successivo decadimento del nucleo eccitato libera un fotone con una energia caratteristica (una «firma») di esattamente 1,2746 milioni di elettronvolt (MeV).

Le stesse esplosioni di supernova che potrebbero aver contribuito a fabbricare gli elementi intermedi e pesanti talvolta lasciano sul luogo dell'esplosione un residuo stellare che ha approssimativamente la massa del Sole, compressa entro un raggio di circa dieci chilometri. La densità nel centro del residuo viene stimata in circa 10^{15} grammi per centimetro cubo, un valore che supera anche la densità all'interno dei nuclei atomici. In effetti la densità è così grande che si pensa che gli elettroni atomici siano stati «conficcati» entro i protoni, annullando la loro carica positiva e trasformandoli quindi in neutroni.

La giovinezza di uno di questi oggetti è assai diversa dalla sua età matura, per quel che se ne sa dalle illazioni fatte dagli astrofisici sulla Terra. Le stelle di neutroni giovani che ruotano rapidamente sono pulsar a radioonde: sorgenti pulsanti di radiazione elettromagnetica a frequenze radio, scoperte per la prima volta negli anni sessanta. La sorgente di energia per le emissioni dei pulsar sembra essere una continua perdita della energia cinetica di rotazione della stella di neutroni, perché per ciascun pulsar conosciuto il periodo misurato fra due impulsi successivi appare in lento aumento.

Due processi non nucleari che generano righe gamma sono schematizzati in questa figura. Nell'annichilazione elettrone-positone *(a sinistra)* un positone incontra un elettrone, la sua antiparticella, e la massa di entrambi viene trasformata in energia elettromagnetica, sotto forma di due fotoni gamma, ciascuno con un'energia di 0,511 MeV. Nell'emissione elettronica di ciclotrone *(a destra)* un campo magnetico intenso e uniforme (rappresentato da un insieme di linee di flusso molto addensate e parallele) sovrappone un moto circolare alla traiettoria di un elettrone. La frequenza di rivoluzione corrisponde alla frequenza (e quindi all'energia) dei fotoni emessi dall'elettrone. Se la frequenza è abbastanza alta i fotoni sono raggi gamma. Questo potrebbe accadere, in particolare, vicino ai poli magnetici di una stella di neutroni.

Il meccanismo effettivo mediante il quale l'energia di rotazione si trasforma nella radiazione osservata rimane un mistero. Ci sono molte teorie in lotta. In generale, tuttavia, si postula che la stella di neutroni possegga un campo magnetico incorporato, in quanto si pensa che il campo della stella genitrice si conservi attraverso tutti gli eventi che portano alla creazione della stella residua. Siccome il residuo è fortemente compresso, lo stesso avviene del campo magnetico, che pertanto raggiunge intensità enormi: si stima che arrivi a 10^{13} gauss (l'intensità del campo magnetico terrestre è minore di un gauss). La stella residua ruota, e il campo magnetico ruota anch'esso. Il risultato è che tale campo rotante induce presso la superficie della sfera un campo elettrico, che «strappa» dallo strato superficiale particelle cariche e le accelera fino a energie elevate. Le traiettorie delle particelle vengono curvate dal campo magnetico stesso, e la curvatura delle traiettorie provoca l'irraggiamento dei fotoni.

Il carattere pulsante della radiazione osservata dalla Terra viene spiegato con l'ipotesi che i due poli del campo magnetico non siano allineati con l'asse di rotazione della stella di neutroni. In questo modo quando la stella ruota ciascuno dei due poli del campo (nelle cui vicinanze l'intensità del campo è particolarmente elevata) alternativamente compare e scompare alla vista. La radiazione emessa viene perciò osservata a impulsi, con quello che viene chiamato «effetto faro».

Per quale ragione tutto ciò è importante per l'astronomia di righe gamma? Peter A. Sturrock, della Stanford University, è stato il primo a precisare una motivazione. Nella maggior parte dei modelli teorici accettati, nel campo magnetico della stella di neutroni vengono creati intensi flussi di positoni. Essi vengono prodotti con un meccanismo curioso. Il campo magnetico è così intenso che quando esso incurva la traiettoria di una particella carica di alta energia la particella emette fotoni gamma. Se l'energia di uno di questi fotoni è maggiore di 1,022 MeV, quando il fotone «urta» il campo magnetico può creare una coppia elettrone-positone (l'energia del fotone deve essere maggiore di 1,022 MeV perché la massa di un elettrone o di un positone corrisponde, applicando la formula $E = mc^2$, a una energia di 0,511 MeV). Quando un positone generato in questo modo incontra un altro elettrone, la coppia si annichila a sua volta, dando origine a fotoni gamma che formano una riga da 0,511 MeV. L'osservazione di questa radiazione strettamente accordata dovrebbe consentire agli astronomi di assegnare dei limiti per la produzione di positoni nei pulsar, e quindi di porre un vincolo importante ai modelli accettabili per i pulsar.

In un milione d'anni o giù di lì anche la stella di neutroni in più rapida rotazione finisce per arrestarsi. La sua esistenza come radiopulsar è terminata, ma anche la sua vecchiaia può essere interessante. Nella nostra galassia sono state individuate centinaia di sorgenti puntiformi di raggi X. Si pensa che la maggior parte di esse siano vecchie stelle di neutroni che stanno aggregandosi materia da stelle compagne in sistemi binari. Mentre la materia aggregata entra vorticosamente nel campo gravitazionale della stella di neutroni, viene accelerata a energie elevate e forma un caldissimo disco di accrescimento. Gli urti fra le particelle all'interno del disco generano l'esteso spettro di raggi X che viene osservato dalla Terra.

Si pensa che la materia in aggregazione sia fortemente ionizzata, e che sia costituita principalmente da protoni ed elettroni di alta energia. Perciò dovrebbe essere possibile osservare anche righe gamma, perché i protoni di alta energia che incidono sulla superficie della stella potrebbero eccitare i suoi nuclei, e il decadimento dei nuclei al loro stato fondamentale genererebbe fotoni gamma. Dalle energie gamma caratteristiche si potrebbe inferire la natura della superficie di una stella di neutroni vecchia.

Una pietra angolare ben salda della teoria della relatività asserisce che un fotone perde energia mentre si arrampica per uscire da un campo gravitazionale, e che la sua lunghezza d'onda cresce corrispondentemente; nel caso della luce il fotone diventa più rosso. Lo spostamento di frequenza è proporzionale alla variazione di energia potenziale gravitazionale, che per la fuga dalla superficie di una stella è proporzionale alla massa della stella divisa per il suo raggio. Poiché la massa di una stella di neutroni è grande, mentre il suo raggio è piccolo, il potenziale gravitazionale è enorme ed enorme è anche la diminuzione di energia dei fotoni che sfuggono dalla superficie. Ci si aspetta pertanto che tutte le righe gamma siano spostate verso energie più basse (e lunghezze d'onda maggiori) per quantità molto grandi. Il parametro di spostamento verso il rosso relativo alla superficie viene indicato con Z_s. Per ogni riga spettrale assegnata esso è dato dalla differenza fra l'energia per fotone osservata e l'energia per fotone che verrebbe misurata qualora la radiazione fosse emessa sulla Terra, espressa come frazione dell'energia osservata.

Interviene ora una seconda linea di discussione, che è stata sviluppata indipendentemente da Kenneth Brecher e da Richard Bowers. Molte stelle di neutroni percorrono orbite, in un sistema binario, che le portano a essere eclissate periodicamente dalla stella compagna. Per ciascuna di queste stelle le interruzioni nella radiazione X osservata sulla Terra forniscono il periodo orbitale e la durata dell'eclisse.

Contemporaneamente, il moto orbitale della stella compagna si svolge alternativamente in direzione della Terra e in direzione opposta, con il corrispondente spostamento Doppler nella sua emissione luminosa. Per giunta, l'intensità dell'emissione luminosa è modulata per effetto delle variazioni della stella compagna sotto l'azione delle forze di marea nel campo gravitazionale della stella di neutroni. Considerate nel loro insieme, tutte queste varie osservazioni permettono di ricavare i parametri dell'orbita della stella di neutroni, e quando si conosce l'orbita si può calcolare la massa della stella.

Il secondo elemento importante della discussione è il fatto che le stelle di neutroni in realtà potrebbero essere costituite in gran parte da particelle subatomiche più esotiche dei neutroni. Nella letteratura scientifica sono state esaminate in dettaglio molte possibilità diverse, inclusi anche i quark, e molti teorici hanno sottolineato il fatto che composizioni diverse portano a raggi (R) diversi per la stessa massa (M). Per ciascuna stella di neutroni una determinazione di M e una misura di Z_S, che è proporzionale a M/R, fornisce una determinazione di R. Ciò limiterà fortemente le proposte sulla composizione della stella. L'astronomia di righe gamma conduce pertanto direttamente alla fisica della materia superdensa.

A questo proposito può essere significativo che recentemente sia stata identificata una nuova categoria di oggetti astronomici. Questi oggetti emettono flussi enormi di radiazione gamma per durate di alcuni secondi o minuti, dopo di che l'emissione si arresta. Per ora non se ne sa molto; in realtà non si sa ancora nemmeno se essi appartengono alla nostra galassia o sono all'esterno di essa. L'ipotesi più credibile è che gli impulsi di raggi gamma siano causati da eventi di alta energia sulla superficie delle stelle di neutroni o vicino ad esse. Questi fenomeni impulsivi potrebbero verificarsi, per esempio, qualora l'acquisizione di materia da parte di una stella di neutroni a spese della stella compagna in un sistema binario avvenisse episodicamente.

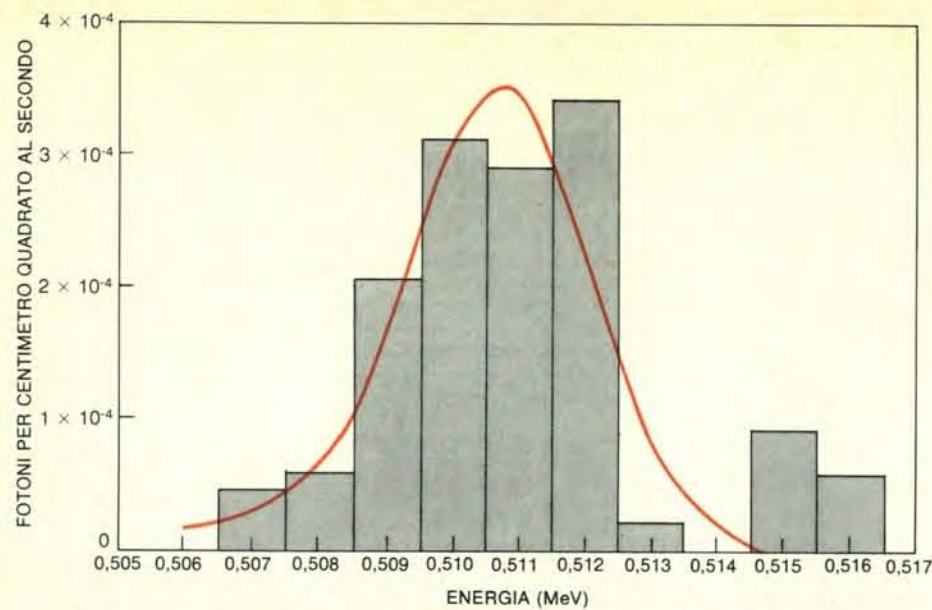
Il 10 giugno 1974 un gruppo guidato da Allan S. Jacobson del Jet Propulsion Laboratory del California Institute of Technology fece puntare un telescopio a righe gamma a grande apertura, portato da un pallone, nella direzione opposta al centro della Galassia, e osservò un impulso di raggi gamma della durata di venti minuti. Il gruppo lo interpretò come un impulso di origine astrofisica. Nello spettro di questo impulso sembrano apparire alcune strutture a righe. È stato proposto che esse rappresentino diversi meccanismi comuni che producono radiazioni gamma, con la particolarità che tutte le righe sono spostate verso il rosso nello stesso rapporto. Per esempio, una struttura a forma di riga attorno a 0,4 MeV viene interpretata come una riga da 0,511 MeV di annichilazione elettrone-positone spostata verso il rosso per uno Z di 0,28. Ci si troverebbe in difficoltà a trovare in quella regione dello spettro elettromagnetico un'altra reazione nucleare che possa essere responsabile per la produzione di quella riga. D'altra parte, il valore di 0,28 per il parametro di spostamento superficiale di una stella di neutroni si accorda bene con i calcoli teorici.

Una osservazione indipendente da parte del nostro gruppo rivelò successivamente una possibile riga a 0,4 MeV, quando il nostro strumento su pallone fu puntato nella stessa direzione. L'apertura di campo del nostro strumento era grande, ma comprendeva la nebulosa del Granchio con la stella di neutroni nel centro. Più recentemente Bonnard J. Teegarden e Thomas L. Cline del Goddard Space Flight Center della NASA hanno pubblicato la prima misura dello spettro di un impulso di raggi gamma da parte di uno strumento a elevata risoluzione portato da un pallone. Anch'essi trovarono qualche indicazione di righe gamma spostate verso il rosso per effetto gravitazionale. Sembra ormai del tutto possibile che si abbiano in mano le prime misure di spostamenti verso il rosso gravitazionali dovuti a una stella di neutroni.

Gli elettroni che incidono sulla superficie di una stella di neutroni offrono la possibilità di un'ulteriore osservazione, a causa del movimento circolare di ciclotrone che possiedono nel campo magnetico della stella. Dove il campo è molto intenso e uniforme, nelle vicinanze dei due poli magnetici della stella, anche la radiazione di ciclotrone potrebbe essere costituita da righe gamma. Questa possibilità fu suggerita per la prima volta dall'astrofisico russo Rashid A. Sunyaev e dai suoi colleghi. L'osservazione di queste righe dovute alla radiazione di ciclotrone equivarrebbe a una misura della intensità

Lo spettro di raggi gamma di Ercole X-1, un'importante sorgente di raggi X, contiene due picchi di raggi gamma, o strutture a riga, a 55 keV (55 000 elettronvolt) e a 110 keV. La riga a 55 keV viene interpretata come radiazione di ciclotrone emessa dagli elettroni nell'intenso campo magnetico di una stella di neutroni, il residuo superdenso di un'esplosione stellare. La riga a 110 keV viene interpretata come un'armonica all'energia doppia di quella dell'emissione di ciclotrone. I dati sono stati raccolti il 3 maggio 1976 da uno strumento trasportato su un pallone lanciato vicino a Palestine, Texas, da un gruppo tedesco dell'Istituto Max Planck e dell'Università di Tübingen, guidato da Joachim Trümper. Il resto dello spettro, che include alla sinistra *(curva tratteggiata)* dati ricavati con uno strumento su satellite, è considerato come la parte superiore di uno spettro continuo di raggi X emessi in urti fra atomi nel disco di materia che circonda la stella di neutroni.

Un impulso di raggi gamma proveniente dalla direzione del cielo opposta a quella del centro galattico fu registrato il 20 giugno 1974 e durò per 20 minuti. Qui sono riprodotte tre diverse regioni di energia del suo spettro. Quattro strutture a riga si elevavano spettacolarmente al di sopra della radiazione di fondo quando il telescopio puntava in quella direzione. Due di esse, indicate con *a* e *b*, potrebbero rappresentare la creazione di fotoni gamma nella reazione che unisce un protone e un neutrone, per formare un nucleo di deuterio. Lo spostamento della struttura *a* a un'energia minore di quella di *b* si spiega ammettendo che i fotoni che costituiscono *a* siano stati prodotti vicino alla superficie di una stella di neutroni e abbiano perso energia sfuggendo dal potente campo gravitazionale della stella. La struttura *c* è attribuita alla annichilazione elettrone-positone, con l'energia dei fotoni anche in questo caso minore del suo valore normale di 0,511 MeV. La struttura *d* è attribuita alla cattura di neutroni da parte di nuclei di ferro 56, che trasformano questi nuclei nello stato eccitato del ferro 57, che decade al suo stato fondamentale con l'emissione di un fotone. Anche in questo caso l'energia dei fotoni osservati risulta minore di quella normale. I dati sono stati raccolti da un gruppo guidato da Allan S. Jacobson presso il Jet Propulsion Laboratory del California Institute of Technology.

Una occasionale acquisizione di gas da parte di una stella di neutroni a spese della stella compagna in un sistema binario può aver dato origine all'impulso di raggi gamma il cui spettro è riprodotto qui in alto. In questa ipotesi, gli urti del gas entrante con il gas del disco che circonda la stella di neutroni danno origine a neutroni e positoni. Gli eventi successivi, che comprendono la cattura di neutroni dall'idrogeno e dal ferro 56 e l'annichilazione mutua di elettroni e positoni, creano i fotoni che compongono l'impulso. I fotoni emessi dalla superficie della stella di neutroni o vicino a essa perdono energia allontanandosi nel campo gravitazionale della stella. Le righe gamma che compaiono nei grafici in alto sono contrassegnate con le lettere da *a* a *d*. La figura è un adattamento da un'illustrazione di R. E. Lingenfelter, J. C. Higdon e R. Ramaty.

La radiazione gamma proveniente dal centro galattico (o almeno da una regione vicina al centro galattico) comprende una struttura a riga attorno a 0,511 MeV. L'intensità della riga corrisponde all'annichilazione di 10^{43} positoni al secondo. Il meccanismo con il quale vengono prodotti i positoni non è ancora noto. Il grafico riassume i dati accumulati dal telescopio a raggi gamma degli autori durante un volo su pallone al di sopra di Alice Springs in Australia nel novembre del 1977.

del campo. Sarebbe anche una prova dell'effettiva esistenza di quegli enormi campi magnetici.

Nel maggio del 1976 un gruppo dell'Istituto Max Planck per la fisica e l'astrofisica di Garching nella Germania Occidentale, in collaborazione con un gruppo dell'Università di Tübingen, entrambi sotto la guida di Joachim Trümper, lanciò un pallone sopra Palestine, nel Texas. Il pallone trasportava uno strumento con il quale essi ottennero uno spettro della radiazione gamma di bassa energia emessa dal pulsar a raggi X Ercole X-1, che è un pulsar a rotazione lenta. I fotoni a raggi gamma di energie molto basse formavano uno spettro continuo, senza righe, che i ricercatori interpretarono come il limite superiore dello spettro di raggi X emesso da un disco di accrescimento che circonda il pulsar. Una struttura a forma di riga appariva, tuttavia, a una energia di 55 keV (55 000 elettronvolt). Una seconda struttura a riga si trovava a 110 keV. Entrambe arrivavano al rivelatore a intervalli di 1,24 secondi, che è il periodo del pulsar. Si pensa che esse siano una riga di radiazione di ciclotrone da elettroni e una armonica di frequenza doppia. L'esistenza della configurazione a 55 keV è stata confermata con un secondo lancio di palloni, e nuovamente l'anno scorso mediante uno strumento a bordo del satellite *HEAO-1* della NASA (il primo High Energy Astronomical Observatory, Osservatorio astronomico ad alta energia). Se si tratta realmente di una riga di ciclotrone, l'osservazione effettuata dal gruppo di Trümper rende possibile per la prima volta la misura del campo magnetico di una stella di neutroni. Se questo è vero, l'intensità del campo di Ercole X-1 è approssimativamente pari a 5,3 × 10^{12} gauss.

Una radiazione diffusa di righe gamma permea tutta la nostra galassia, e si pensa perciò che emani dal piano della Via Lattea. Il suo studio richiederà telescopi a raggi gamma più sensibili di quelli attualmente in funzione. Questa radiazione si genera perché particelle della radiazione cosmica (per la maggior parte nuclei di alta energia) urtano contro granelli di polvere, molecole e atomi presenti nello spazio interstellare ed eccitano i loro nuclei a livelli energetici al di sopra dello stato fondamentale. Il decadimento di tali nuclei dà poi origine ai raggi gamma. Siccome il granello di polvere, la molecola o l'atomo urtati da una particella veloce rinculano, il nucleo si trova in movimento quando il fotone viene emesso. Il risultato è uno spostamento Doppler per l'energia del fotone. Si tratta di uno spostamento verso una energia più alta se il nucleo si muove nella direzione del sistema solare, verso una energia più bassa se il nucleo se ne allontana. Conseguentemente, per un gruppo di nuclei che si muovono a caso le righe gamma emesse risulteranno allargate.

L'allargamento è proporzionale alla velocità dei nuclei. Perciò Reuven Ramaty, Benzion Kozlovsky, Richard E. Lingenfelter e altri hanno ipotizzato che ogni riga gamma possa essere una combinazione di tre contributi distinti. Per ogni riga una componente molto stretta deriverebbe dall'emissione di fotoni dai nuclei degli atomi appartenenti a granelli di polvere, che hanno una elevata probabilità di perdere la loro velocità di rinculo all'interno del granello prima di emettere la radiazione. La componente stretta delle righe potrebbe perciò fornire una buona possibilità per determinare la composizione e la distribuzione della polvere interstellare.

Una componente alquanto più diffusa di ciascuna delle righe dovrebbe originarsi dall'emissione di fotoni da parte dei nuclei di atomi non confinati entro granelli di polvere. Questa componente dovrebbe perciò fornire una buona opportunità per determinare la composizione e la distribuzione del gas interstellare. Naturalmente anche altre forme di astronomia possono servire al medesimo scopo. Nel caso dell'astronomia ottica, tuttavia, la presenza della polvere interstellare limita le osservazioni dirette verso il centro della Galassia a una distanza di soli 10 000-13 000 anni luce, mentre il raggio della Galassia è più di 30 000 anni luce. La materia interstellare è molto trasparente per i raggi gamma. Inoltre, occorre tener presente che i fotoni ottici e X sono emessi dagli elettroni atomici, così che la struttura delle righe di emissione viene complicata dallo stato chimico della materia, per esempio dal grado di ionizzazione. Il fatto che i fotoni gamma sono invece emessi dai nuclei degli atomi garantisce che la loro struttura spettrale è indipendente dallo stato chimico della materia.

Infine una componente diffusa di ciascuna riga dovrebbe derivare dall'emissione di fotoni da parte dei nuclei dei raggi cosmici, le cui velocità ammontano a una considerevole percentuale della velocità della luce. La componente diffusa offre così la possibilità di studiare anche i raggi cosmici, i cui urti con la materia interstellare sono considerati la causa della presenza di regioni calde entro il gas interstellare. Si sospetta anche che gli urti dei raggi cosmici, spaccando i nuclei degli atomi interstellari, contribuiscano alla creazione di elementi leggeri come il litio e il berillio (si possono anche ipotizzare reazioni di fusione nucleare che creino questi elementi nell'interno delle stelle, ma esistono altre reazioni che li consumerebbero). La difficoltà della rivelazione diretta dei raggi cosmici, così come arrivano sulla superficie della Terra, consiste nel fatto che i raggi interagiscono con i campi magnetici del sistema solare, così che il campione che arriva sulla Terra potrebbe non essere rappresentativo di quello esistente nello spazio interstellare. La rivelazione delle loro righe gamma caratteristica eliminerebbe il problema.

Altri oggetti importanti per l'astronomia di righe gamma sono le galassie e le loro regioni centrali. I quasar, per esempio, sono considerati come galassie giovani, enormemente distanti dalla Terra e purtuttavia notevolmente luminose. Per la loro interpretazione vengono usati due tipi di modelli. I modelli del primo tipo ipotizzano che al centro del quasar avvengano numerose esplosioni di supernova. I modelli del secondo tipo suppongono invece che le stelle vicine al centro del quasar siano lacerate dal campo gravitazionale di un enorme buco nero e che la materia stellare formi un disco di accrescimento caldo attorno a esso. Nel secondo caso ci sarebbe la speranza di osservare la riga gamma di emissione dagli urti all'interno del disco, con tutte le righe fortemente spostate verso il rosso e anche allargate a causa del campo gravitazionale del buco nero.

Anche il centro della nostra galassia è una regione misteriosa, in parte perché i telescopi ottici non riescono a rivelarne la struttura. Nel novembre 1977 e poi ancora nell'aprile 1979, il nostro gruppo ha lanciato il suo strumento con un pallone al di sopra di Alice Springs in Australia, nel tentativo di rivelare righe gamma provenienti dalla direzione del centro galattico. In un esperimento precedente un gruppo della Rice University, guidato da Robert C. Haymes, aveva ottenuto un indizio allettante dell'emissione di una riga attorno a 0,5 MeV. Nel nostro lancio del 1977 trovammo una riga esattamente a 0,511 MeV, nel mezzo di uno spettro continuo di raggi gamma. Avemmo conferma dell'esistenza della riga nel 1979. Essa corrisponde all'annichilazione di 10^{43} positoni per secondo nel centro della Galassia o in prossimità di esso. L'esistenza di questa riga proveniente dal centro della Galassia era stata postulata due decenni fa da Philip Morrison del Massachusetts Institute of Technology, e più tardi da Floyd W. Stecker del Goddard Space Flight Center e da uno di noi (Leventhal).

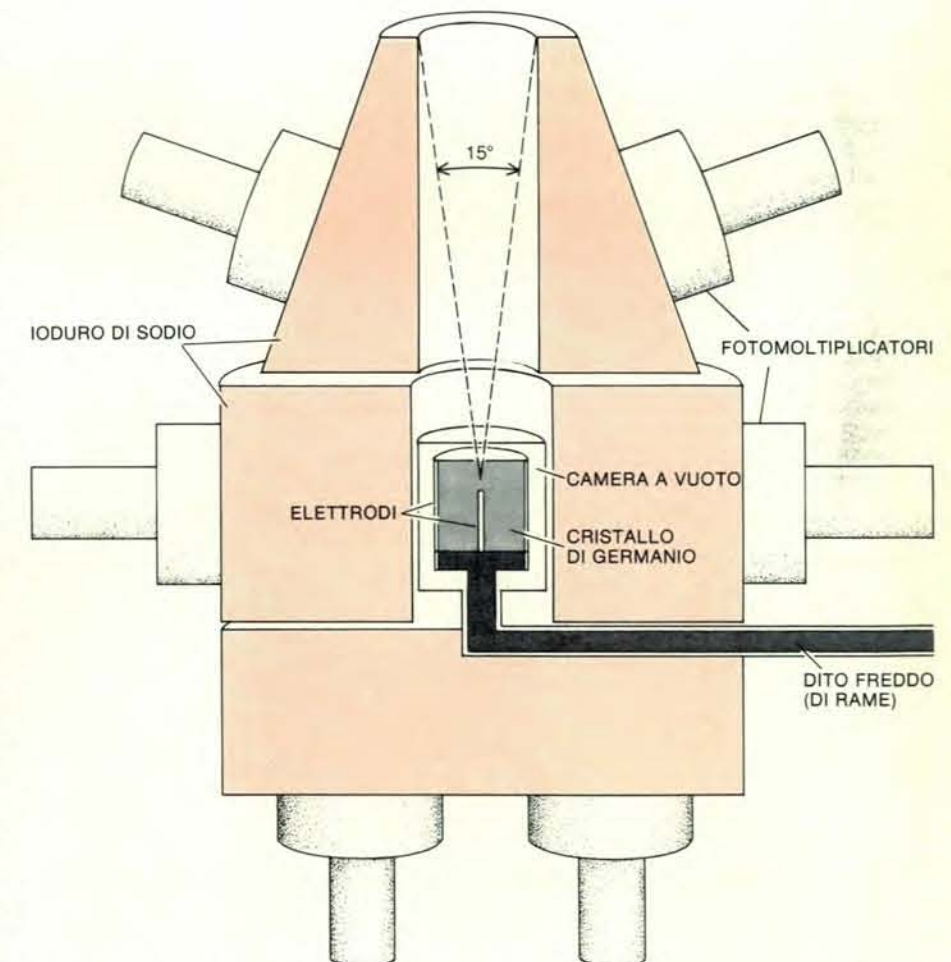
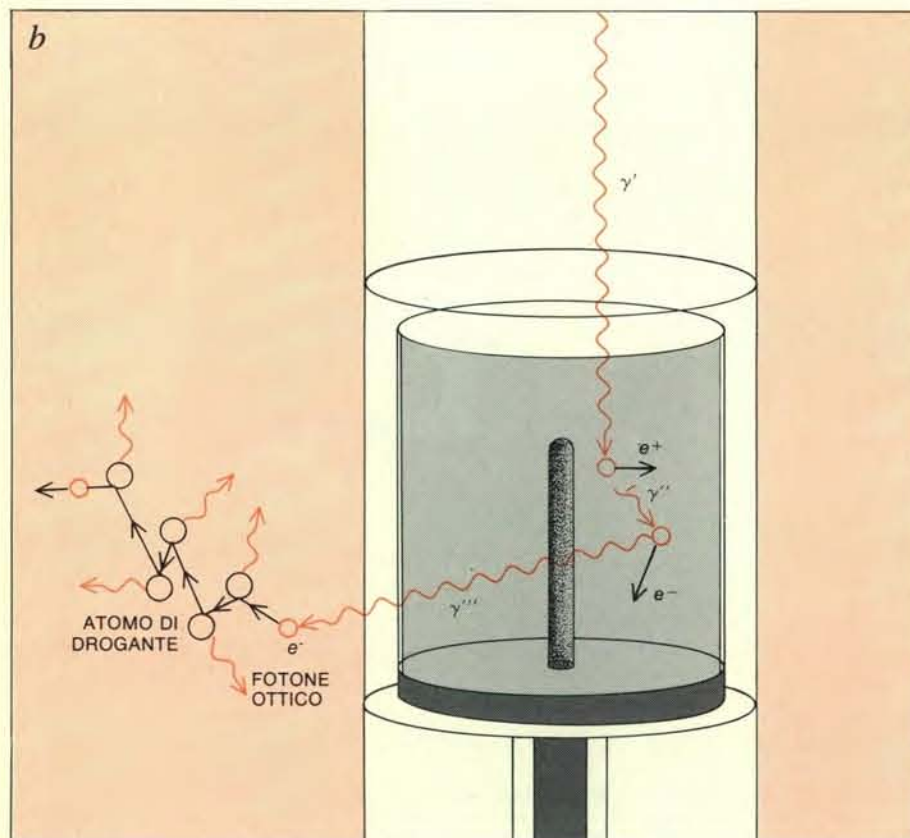
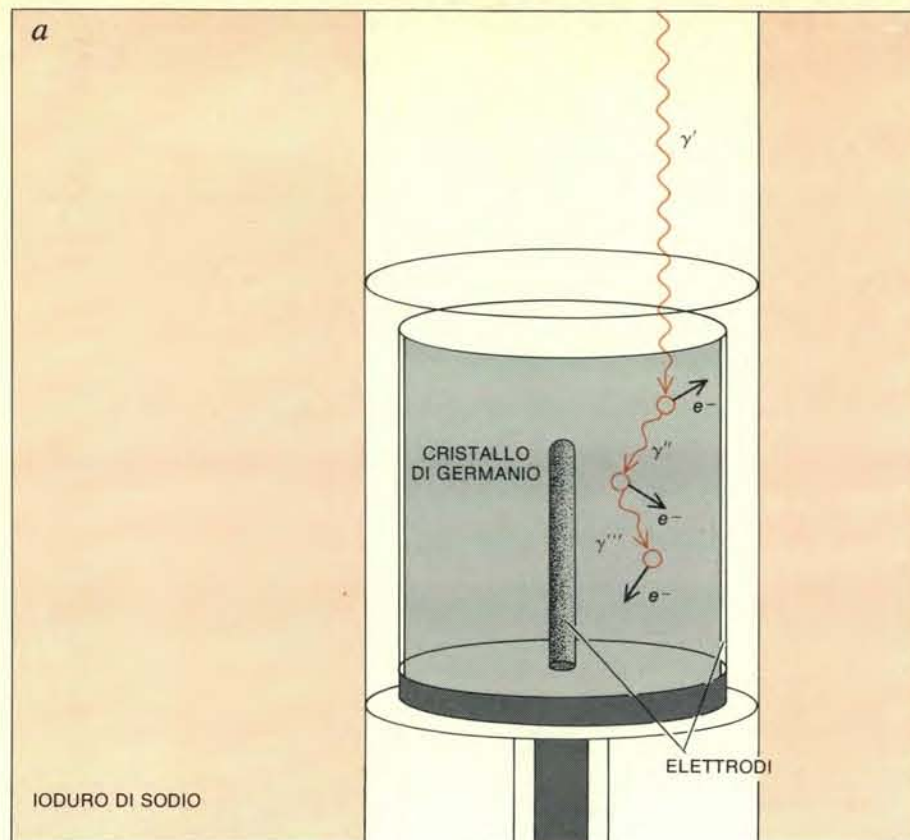
La letteratura scientifica è ricca di suggerimenti circa l'origine dei positoni. In quella che è probabilmente l'ipotesi più accreditata i positoni vengono considerati come un sottoprodotto delle supernove. L'idea è che il decadimento dei nuclei radioattivi prodotti nelle esplosioni stellari produce i positoni, che si allontanano con energie di alcune centinaia di migliaia di elettronvolt. Possono trascorrere migliaia di anni prima che un positone urti contro un elettrone e si annichili. Perciò la riga osservata dalla Terra può rappresentare esplosioni di supernova avvenute su intervalli di tempo molto lunghi. Il fatto che le stelle sono molto numerose nel centro galattico gioca a favore di questa ipotesi.

Una seconda possibilità è che i raggi cosmici urtino contro la materia interstellare e spacchino i suoi nuclei trasformandoli in nuclei radioattivi, il cui decadimento genera i positoni. Gli urti possono anche generare mesoni pi, il cui decadimento dà origine ai positoni. Calcoli recenti indicano tuttavia che i valori correnti per il flusso di raggi cosmici e la densità della materia interstellare sono di due ordini di grandezza troppo piccoli per dar conto dell'intensità della riga da 0,511 MeV da noi osservata. Quell'intensità è veramente grande: è come se la luminosità del Sole venisse moltiplicata per 10 000 e concentrata in un'unica frequenza dello spettro elettromagnetico.

Una terza possibilità è che il centro galattico ospiti un gran numero di radiopulsar, che proiettano grossi fasci di positoni nello spazio interstellare. Altre possibilità sono al momento attuale meno credibili. C'è un'ipotesi che propone l'esistenza di un grande buco nero al centro della Galassia, circondato da un grande disco di accrescimento. Un'altra suppone che piccoli buchi neri siano stati creati in gran numero nell'universo primordiale e che essi stiano, per così dire, evaporando attraverso processi quantistici, irraggiando la loro energia principalmente sotto forma di coppie elettrone-positone. Forse l'ipotesi meno plausibile è che nel centro della nostra Galassia esista una grossa massa di antimateria.

Il punto importante, tuttavia, è che il meccanismo con il quale i positoni sono generati nel centro della Galassia può essere individuato mediante osservazioni che potranno essere eseguite nel prossimo decennio, forse entro i prossimi cinque anni. Esso verrà determinato per mezzo di mappe dettagliate della regione di emissione. Se tale zona è una sorgente puntiforme, l'emettitore probabile è un buco nero. Se compaiono dei picchi vicino a radiopulsar, questi ultimi sono le sorgenti probabili. Inoltre, molte di queste ipotesi richiedono l'esistenza di altre righe oltre a quella da 0,511 MeV, così che l'osservazione di tali righe limiterà il numero delle possibilità.

Il fattore cruciale, in merito a un telescopio per righe gamma, è che i fotoni gamma non possono essere riflessi mediante specchi. Essi penetrano nella materia e tendono a perdere energia tramite diffusioni o deviazioni in urti contro elettroni. Questo tipo di diffusioni domina il progetto del telescopio. Al centro del telescopio lanciato da noi vi è un cristallo di germanio con un volume di 130 centimetri cubi. All'interno del cristallo un fotone gamma incidente urta diversi elettroni uno dopo l'altro. Il rinculo degli elettroni genera una piccola corrente, di intensità proporzionale all'energia del raggio gamma. I fotoni gamma possono così essere rivelati uno alla volta.

Attorno al rivelatore centrale sono disposti circa 18 chilogrammi di ioduro di sodio, drogato con atomi di tellurio. Anche questo è un materiale sensibile ai rag-

Il telescopio per righe gamma lanciato dagli autori e dai loro collaboratori è costituito da due rivelatori uno dentro l'altro. Al centro vi è un cristallo di germanio, nel quale il fotone gamma incidente viene diffuso dagli elettroni. Il rinculo degli elettroni dà origine a un impulso di corrente elettrica. Il germanio viene mantenuto freddo da un «dito freddo» di rame, raffreddato con azoto liquido; se non fosse così la corrente spontanea del cristallo sommergerebbe la corrente generata da un raggio gamma. Intorno al cristallo centrale sono disposti blocchi di ioduro di sodio drogato con piccole impurezze di atomi di tellurio. Qui un fotone gamma incidente viene assorbito da un elettrone, che successivamente urta contro gli atomi del drogante: la radiazione luminosa emessa dagli atomi eccitati viene poi amplificata mediante fotomoltiplicatori. I dettagli del procedimento per la rivelazione di un raggio gamma sono descritti nella figura della pagina seguente.

Il telescopio a righe gamma degli autori rivela i fotoni gamma a uno a uno. In *a* un fotone incidente (γ') viene diffuso più volte, ma rimane entro il cristallo di germanio centrale. Il rinculo degli elettroni dai quali viene diffuso provoca una corrente elettrica proporzionale all'energia del fotone. In *b* un fotone in arrivo viene diffuso entro lo ioduro di sodio circostante, dove provoca l'emissione di luce. Il fotone cede nel cristallo centrale solo parte della sua energia; la sua rivelazione è vietata dal segnale provocato dal lampo di luce nello ioduro di sodio. L'evento di tipo *b* è più frequente.

gi gamma; l'urto di un fotone gamma con un elettrone provoca il rinculo dell'elettrone, che poi urta con molti atomi del drogante, e gli atomi urtati a loro volta emettono impulsi di luce con fotoni a energie ottiche.

La risoluzione in energia del cristallo di germanio a 1 MeV è all'incirca 2 keV. Ciò significa che fotoni da 1 MeV danno luogo a una riga di larghezza 2 keV. La risoluzione del rivelatore a ioduro di sodio è peggiore per circa un fattore 30. Tuttavia lo ioduro di sodio assolve a tre funzioni importanti. In primo luogo lascia aperta una finestra verso il rivelatore centrale, definendo così il campo visuale del telescopio. In secondo luogo, protegge il rivelatore centrale da un intenso fondo di fotoni gamma generati nell'atmosfera terrestre dai raggi cosmici incidenti dallo spazio. In terzo luogo, gli impulsi di luce prodotti nello ioduro di sodio vengono amplificati mediante fotomoltiplicatori e usati poi «in anticoincidenza» per vietare la rivelazione di quei fotoni gamma che vengono diffusi fuori dal cristallo di germanio e finiscono entro il cristallo di ioduro. Questi fotoni depositano la loro energia in parte nell'uno e in parte nell'altro materiale, così che la corrente nel rivelatore centrale non è proporzionale all'energia effettiva del fotone.

Nel corso di un volo della durata di uno o due giorni il nostro telescopio viene guidato da una stazione telemetrica a terra in modo che punti in una certa direzione. Quando il telescopio è alla ricerca di una sorgente di radiazione gamma, ogni centimetro quadrato della superficie del rivelatore centrale viene colpito mediante un fotone gamma di origine astrofisica ogni parecchi minuti. C'è più di una probabilità su due, tuttavia, che il fotone venga diffuso fuori dal cristallo centrale. La rivelazione del fotone verrà allora «vietata» dalla quasi simultanea produzione di luce nello schermo circostante. Così a poco a poco il rivelatore centrale accumula uno spettro di conteggi di raggi gamma in funzione dell'energia dei fotoni. Alla fine del volo il telescopio viene liberato dal pallone a una quota di circa quaranta chilometri e ricondotto a terra per mezzo di un paracadute.

A noi sembra che gli esperimenti a bordo di palloni continueranno a essere importanti nell'astronomia di righe gamma. In primo luogo, intensi flussi di particelle di alta energia sono presenti nelle fasce di radiazione alle quote alle quali vengono lanciati i satelliti. Le particelle generano un intenso fondo di raggi gamma. D'altra parte, gli esperimenti portati da palloni sono meno costosi e più brevi degli esperimenti portati da satelliti. Dal momento in cui viene avanzata la proposta al momento in cui comincia a produrre risultati, un esperimento su satellite richiede dai cinque ai dieci anni e costa decine di milioni di dollari. Un esperimento su pallone può richiedere meno di due anni e costare meno di un milione di dollari. E finalmente, i palloni costituiscono un eccellente terreno di prova per futuri strumenti da satellite. D'altro canto, i satelliti rimangono in orbita per periodi di tempo piuttosto lunghi e possono pertanto raccogliere quantità di dati molto grandi.

Recentemente è stata costituita una rete di satelliti, capace di stabilire la direzione di origine degli impulsi di raggi gamma con grande precisione mediante triangolazione. Il 5 marzo 1979 questa rete permise di rivelare l'impulso di raggi gamma di energia totale più elevata finora scoperto e di localizzarne l'origine nel centro dei resti di una esplosione di supernova nella Grande Nube di Magellano, una galassia satellite della nostra. Uno spettro dell'impulso fu ricavato da E. P. Mazets e dai suoi colleghi presso l'Istituto fisico-tecnico A. F. Ioffe a Leningrado. Questo spettro comprende una struttura di riga a 0,4 MeV: l'energia di una riga di annichilazione di positoni spostata verso il rosso. La posizione dell'impulso e la presenza della riga suggeriscono con chiarezza che l'origine degli impulsi di raggi gamma sia veramente costituita dalle stelle di neutroni.

Lo scorso autunno la NASA lanciò il satellite *HEAO-3*. Questo trasporta un telescopio per raggi gamma costituito da quattro rivelatori di germanio entro una schermatura di ioduro di sodio. Lo strumento, che è stato costruito da un gruppo di ricerca guidato da Jacobson presso il Jet Propulsion Laboratory, è progettato per funzionare per un anno, scandagliando la maggior parte della volta celeste alla ricerca di righe gamma nette, con una sensibilità massima di 10^{-4} fotoni al secondo per centimetro quadrato di superficie del rivelatore. Lo strumento funziona regolarmente, ma è ancora troppo presto per poter riportare qualche risultato preciso.

La NASA sta ora progettando di lanciare un osservatorio orbitante per raggi gamma nella metà degli anni 1980. Il satellite trasporterà cinque strumenti sensibili ai raggi gamma, dei quali tre saranno telescopi a spettro largo, mentre due ricercheranno righe gamma: un telescopio a nucleo di germanio progettato dal gruppo di Laurence E. Peterson dell'Università della California a San Diego e un telescopio a nucleo di ioduro di sodio progettato dal gruppo di James D. Kurfess del Naval Research Laboratory. Il gruppo di Peterson comprende ricercatori presso i Bell Laboratories, il Goddard Space Flight Center, il Jet Propulsion Laboratory, i Sandia Laboratories e due istituzioni francesi: il Centro di ricerche nucleari di Saclay e il Centro per lo studio dei raggi cosmici di Toulouse. Il gruppo di Kurfess comprende ricercatori della Northwestern University, della Rice University e del Royal Naval College in Gran Bretagna. Gli strumenti dovrebbero scandagliare la maggior parte del cielo durante i due anni e mezzo previsti per l'impresa, e avrebbero una sensibilità di riga di 10^{-5} fotoni al secondo per centimetro quadrato. Con il lancio di quell'osservatorio l'astronomia di righe gamma diventerà maggiorenne.

Intelligenza e afasia

I test hanno dimostrato che gli afasici soffrono di un deficit intellettuale specifico; questi pazienti tuttavia possono recuperare il linguaggio con la rieducazione e svolgere compiti anche complessi purché già appresi

di Anna Basso

L'afasia è un disturbo acquisito del linguaggio conseguente alla lesione di determinate aree cerebrali: le aree del linguaggio. Casi di perdita della parola sono ripetutamente citati nel *Corpus Hippocraticus,* ma non è chiaro se gli ippocratici distinguessero tra i vari tipi di mancanza della parola; per esempio tra afasia e perdita della voce. La storia «scientifica» dell'afasia si fa convenzionalmente risalire al 1861, anno in cui il chirurgo francese Pierre-Paul Broca presentò alla Société d'Anthropologie di Parigi i risultati dell'autopsia di un paziente affetto da afasia, attribuendo la perdita della parola alla lesione della 2ª o 3ª circonvoluzione frontale. Quattro anni dopo, il 15 giugno 1865, sempre davanti alla Société d'Anthropologie, Broca poneva le basi di tutto il problema della dominanza cerebrale collegando la dominanza emisferica per il linguaggio a quella manuale: nei soggetti destrimani le aree del linguaggio sono localizzate nell'emisfero di sinistra che controlla i movimenti della mano destra, nei soggetti mancini nell'emisfero destro che controlla i movimenti della mano sinistra. Broca descrisse l'afasia, da lui chiamata afemia, come la perdita «della facoltà di coordinare i movimenti del linguaggio articolato» e da allora quei pazienti che presentano afasia con difficoltà nella articolazione delle parole vengono chiamati afasici di Broca o, con un termine entrato in uso più recentemente, afasici non fluenti.

Nel 1874 il neurologo tedesco Karl Wernicke descrisse una diversa forma di afasia - che da lui prese il nome di afasia di Wernicke (o afasia fluente) - conseguente alla lesione di una zona situata posteriormente e inferiormente all'area di Broca: l'area di Wernicke nel terzo posteriore della prima circonvoluzione temporale. Wernicke, contrariamente a Broca, pose l'accento sul disturbo di comprensione del linguaggio dal quale fece derivare anche il deficit di espressione e il deficit del linguaggio scritto.

Da allora gli studi sull'afasia sono stati numerosi sia per l'interesse dell'argomento in sé - in quanto colpisce forse la funzione più «umana» che ci sia e cioè il linguaggio - sia per il notevole numero di persone affette da questo disturbo. Nella maggior parte dei casi l'afasia consegue a una lesione vascolare - emboli, malformazioni arterovenose, complicazioni dell'arteriosclerosi, ipossia cerebrale acuta - ma cause frequenti sono anche i traumi e i tumori cerebrali.

Un argomento molto dibattuto per il suo enorme interesse pratico è stato quello del recupero, sia del recupero cosiddetto spontaneo sia di quello conseguente a trattamento riabilitativo. È questo un argomento particolarmente delicato perché non si può decidere di non rieducare certi pazienti al solo scopo di avere un gruppo di controllo che permetta di valutare la reale efficacia di un trattamento riabilitativo. L'anno scorso, valutando i dati raccolti negli ultimi 15 anni, siamo comunque riusciti a dimostrare che un paziente afasico ha maggiori possibilità di recupero in tutti i comportamenti verbali utili alla comunicazione - espressione e comprensione del linguaggio orale e scritto - se viene sottoposto a rieducazione e che queste possibilità sono tanto maggiori quanto più precoce è il trattamento.

L'afasia ha suscitato l'interesse di numerosi studiosi di varie discipline (neurologi, psicologi e linguisti in particolare); l'argomento tuttavia è tale per cui sono ancora molti i punti in discussione. Uno di questi riguarda i rapporti tra deficit intellettuale e afasia; vediamo come sono stati interpretati nel corso di questi ultimi decenni e se è possibile trarre qualche conclusione dagli studi sperimentali effettuati in questo campo.

La scuola dei neoticisti

Il problema dei rapporti tra intelligenza e afasia risale agli albori degli studi sull'afasia, quando ancora non era chiaro cosa si intendesse con questo termine. Il noto medico francese Armand Trousseau, per esempio, riferì nel 1861 il caso di un collega che subì un episodio di afasia durato qualche ora; in quel periodo non era più in grado né di parlare, né di capire quanto leggeva. Trousseau considerava che l'afasia consistesse esclusivamente nella perdita dell'eloquio, mentre per lui l'incapacità di leggere e scrivere non facevano parte del quadro afasico, ma li considerava una perdita intellettuale; riteneva inoltre che l'afasia non potesse regredire rapidamente.

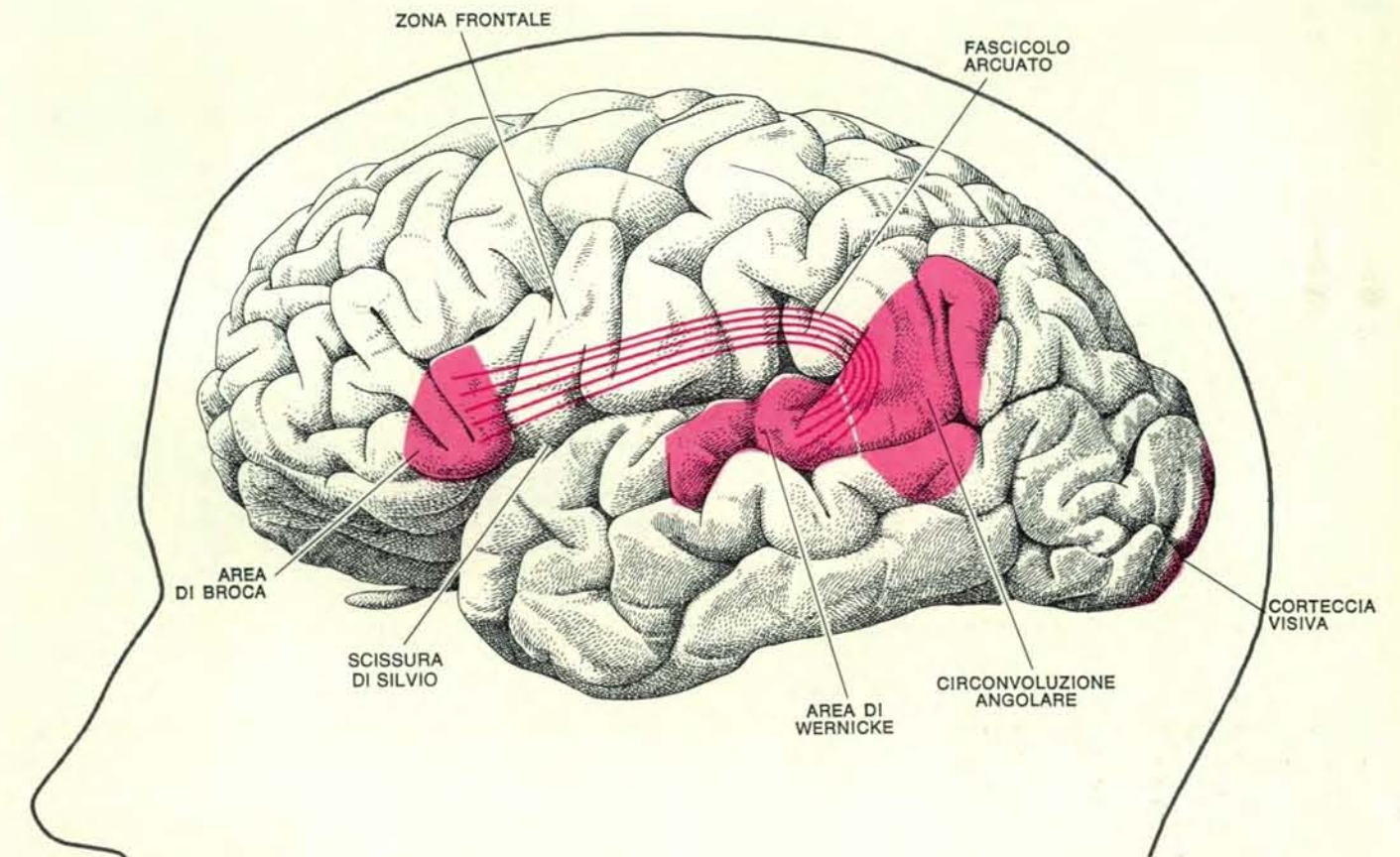
Un'ipotesi ripetutamente formulata nel corso degli studi sull'afasia propone che il deficit intellettuale sia la causa e non la conseguenza dell'alterazione del linguaggio; i sostenitori di questa tesi vengono chiamati «noeticisti» dal greco *nous* = mente. Quale esempio di questa corrente di pensiero possiamo citare Finkelnburg che nel 1870 ipotizzava nell'afasia un deficit più generale e più fondamentale del solo deficit del linguaggio: una difficoltà nell'uso dei segni, una disorganizzazione cioè della funzione simbolica. Questa difficoltà nell'uso dei segni si rifletteva in tutte le manipolazioni delle nozioni concettuali. La perdita della capacità di cogliere e di esprimere dei concetti per mezzo di simboli appresi, non solo linguistici, venne da lui detta asimbolia. Ma è solo nel 1906 con il medico e neurologo francese Pierre Marie che l'afasia viene chiaramente identificata e spiegata come un deficit intellettuale, come una «demenza specializzata». Contro coloro che sostenevano esservi varie forme di afasia ognuna delle quali corrispondente al deficit di un «momento» del linguaggio - espressione, comprensione, articolazione ecc. - e conseguente alla lesione di aree diverse nell'emisfero di sinistra (Bastian, Charcot, Wernicke, ecc.), Pierre Marie disse che l'afasia è una sola: quella descritta da Wernicke (quella descritta da Broca non sarebbe altro che un'afasia di Wernicke più anartria). Nel 1906 l'alternativa era tra un'afasia considerata come un deficit «sensoriale» o un'afasia considerata come un deficit «intellettuale»: non c'era una via di mezzo tra sensazione e intelligenza.

Pierre Marie nega comunque che l'afasia di Wernicke sia un'afasia «sensoriale», in quanto egli ritiene che gli afasici non siano più in grado di intendere non le parole in se stesse, ma la loro combinazione in frasi, neppure se queste vengono espresse gestualmente. Il disturbo è quindi a un livello più alto della semplice sensazione. «Se dovessi dare una definizione di afasia - scrive Pierre Marie - il fatto che mi sforzerei di mettere in luce sarebbe la diminuzione dell'intelligenza.» Sempre secondo Pierre Marie non è però tutta l'intelligenza a essere compromessa, ma solo «il complesso delle cose apprese per via didattica». Pierre Marie oppone al linguaggio, inteso come strumento privo di funzione intellettuale (e alla sua disintegrazione nell'afasia), l'intelligenza che comprende anche il linguaggio considerato come un'attività simbolica concettuale. Dopo Pierre Marie, il maggior sostenitore di un deficit intellettuale come causa (e non conseguenza) del deficit del linguaggio è Kurt Goldstein che, nell'ambito delle teorie della psicologia della forma, ritiene che nell'afasia si abbia una modificazione dei rapporti tra figura e sfondo: la lesione cerebrale smussa la dominanza del processo di figura rispetto al processo di sfondo. Un'indifferenziazione del rapporto figura/sfondo provocherebbe una regressione riducendo il soggetto alla situazione presente; il cerebroleso perderebbe il pensiero categoriale, la capacità cioè di rappresentarsi i rapporti che esistono tra la cosa presente e ciò che questa simbolizza. L'afasico, per esempio, non riuscirebbe a staccarsi dalla situazione concreta: un coltello presentato vicino a una matita è un «temperino», presentato vicino a del pane è un «coltello da pane», ecc.; egli non è più in grado di astrarre dalla situazione attuale per risalire alla categoria «coltelli». Goldstein si chiede se per questi malati tutte le parole non abbiano il valore di nomi propri.

Nel 1962 Eberhard Bay sostiene che il disturbo centrale nell'afasia è un'alterazione nella formazione dei concetti. A sostegno dell'esistenza di un deficit intellettuale negli afasici, Bay porta quale esempio la loro difficoltà a eseguire modelli in creta, il che sarebbe indicativo di un difettoso concetto sottostante. Un elemento comune ai noeticisti, oltre a quello di considerare un deficit intellettuale all'origine del disturbo afasico, è quello di ritenere che l'afasia sia una soltanto (non vi siano, cioè, forme distinte di afasia). Senza comunque voler analizzare quale può essere stato il contributo di questo filone di pensiero nell'interpretazione del deficit afasico, vediamo di capire cosa si intendeva in quegli stessi anni per intelligenza.

La definizione di intelligenza

La psicologia come scienza è molto recente; solo alla fine del secolo scorso essa si stacca decisamente dalla filosofia che da sempre si era interessata ai problemi della psiche. È solo con Herbert Spencer e Hippolyte Taine che si cominciano a distinguere le concezioni filosofiche da quelle pragmatico-sperimentali di intelligenza. Ma cosa si intende per intelligenza?

Nel 1883 lo scienziato inglese Francis Galton nel suo laboratorio al Kensington Museum di Londra sottoponeva a quelli che lui chiamava test mentali i visitatori del museo; misurava, tra le altre cose, la statura, i tempi di reazione e l'acuità visiva. Negli ultimi anni del secolo cominciava però a farsi strada in Europa con Alfred Binet, C. Spearman e altri,

Si ritiene che nel cervello umano le zone primarie del linguaggio siano situate nell'emisfero sinistro, perché solo raramente una lesione all'emisfero destro provoca disturbi del linguaggio. L'area di Broca, adiacente alla regione della corteccia motoria che controlla il movimento dei muscoli delle labbra, della mascella, della lingua, del palato molle e delle corde vocali, a quanto sembra contiene i programmi per il coordinamento di questi muscoli nel linguaggio. Un danno all'area di Broca provoca un'articolazione lenta e faticosa del linguaggio (afasia non fluente). L'area di Wernicke è situata fra la circonvoluzione di Heschl, la stazione che riceve originariamente gli stimoli uditivi, e la circonvoluzione angolare, che funge da stazione intermedia fra le regioni uditive e quelle visive. Quando è danneggiata l'area di Wernicke il discorso è fluente, ma scarso di contenuto, e in genere manca la comprensione (afasia fluente). L'area di Wernicke e quella di Broca sono unite da un fascio di fibre chiamato fascicolo arcuato. Quando questo fascicolo è danneggiato il discorso è fluente, ma anormale (afasia di conduzione).

	SOGGETTI NON RIEDUCATI (119)											
	AFASIA NON FLUENTE						AFASIA FLUENTE					
	GRAVI (34)			MODERATI (19)			GRAVI (25)			MODERATI (41)		
TEMPO (MESI)	<2	2-6	>6	<2	2-6	>6	<2	2-6	>6	<2	2-6	>6
MIGLIORATI	1	0	0	6	1	0	3	0	0	11	2	1
INVARIATI	10	13	10	3	5	4	10	7	5	19	4	4

In queste tabelle sono riassunti i dati raccolti a proposito della espressione spontanea di 281 pazienti esaminati sei mesi dopo la loro prima visita eseguita presso il Servizio di rieducazione dell'afasia della Clinica neurologica dell'Università di Milano. Sia i soggetti rieducati che quelli non rieducati sono stati classificati in base ai due principali tipi di afasia: non fluente e fluente; in base alla gravità della malattia e al tempo intercorso tra il suo manifestarsi e la prima visita. L'efficacia della rieducazione risulta evidente dal confronto dei due gruppi così come si può facilmente notare che i progressi maggiori si ottengono nei soggetti presi in cura più precocemente. Anche se si sono osservati miglioramenti spontanei, la rieducazione si dimostra sempre consigliabile in quanto riesce ad aumentare le probabilità di ripresa del paziente.

	SOGGETTI NON RIEDUCATI (162)											
	AFASIA NON FLUENTE						AFASIA FLUENTE					
	GRAVI (23)			MODERATI (54)			GRAVI (43)			MODERATI (42)		
TEMPO (MESI)	<2	2-6	>6	<2	2-6	>6	<2	2-6	>6	<2	2-6	>6
MIGLIORATI	2	1	0	16	10	6	12	5	1	14	5	3
INVARIATI	7	7	6	2	10	10	15	8	2	6	8	6

l'idea che l'intelligenza di una persona fosse una qualità globale del suo comportamento generale variabile da persona a persona e che non fosse legata a specifiche abilità (come l'acuità visiva, i tempi di reazione ecc.).

È però solo nel 1905 con A. Binet e T. Simon, (l'anno prima che Pierre Marie dicesse che l'afasia era un disturbo dell'intelligenza) che si ha la prima grande rivoluzione nella comprensione e valutazione dell'intelligenza. Binet, laureato in legge, si era però sempre interessato di psicologia, dapprima per conto proprio, poi con J. M. Charcot alla Salpêtrière e infine con H. E. Beaunis alla Sorbona. Egli era lentamente giunto a concepire l'intelligenza come una caratteristica del comportamento umano globale che non poteva essere suddivisa in facoltà specifiche, intendendo la parola comportamento nel senso più vasto del termine e cioè come modo di agire di una persona in una data situazione. Nel novembre 1904 Binet fu nominato membro della Commissione per lo studio dei bambini ritardati che aveva il compito di mettere a punto delle prove che permettessero di individuare quei bambini che non erano in grado di seguire un regolare corso di studi. Con l'aiuto di Thomas Simon, Binet in meno di un anno preparò il primo test di intelligenza che - con varie e successive revisioni - resta tuttora quello più usato. Egli affrontò subito il problema in modo pratico; avendo constatato empiricamente che vi sono delle differenze di comportamento a seconda dell'età, costruì un test eterogeneo con prove diverse per ogni anno di età; erano considerate valide e fecero parte della versione finale del test tutte quelle prove che venivano superate da una sempre maggior percentuale di bambini con il crescere dell'età ed erano attribuite all'età in cui erano superate dal 60-70 per cento dei bambini; le prove che venivano risolte con uguale facilità da bambini di età diversa venivano scartate. Nel 1908 - anno del grande dibattito tra Pierre Marie sostenitore di un deficit intellettuale nella afasia e J. Dejerine che negava che gli afasici avessero un deficit intellettuale - Binet e Simon introdussero il concetto di età mentale. L'età mentale di un bambino si calcola contando quante prove del test il bambino ha risolto correttamente; nell'ultima versione del test di Binet (quella curata da L. M. Terman e Merrill nel 1937) a ogni prova corrisponde un valore di 2 mesi. Conoscendo il numero di prove risolte si può calcolare l'età mentale del bambino e confrontarla con la sua età cronologica; troviamo così bambini che hanno età mentali corrispondenti o meno alla loro età cronologica. Il concetto di quoziente intellettuale o d'intelligenza (Q. I.) è stato introdotto pochi anni dopo, nel 1912, da William Stern e si ottiene moltiplicando per 100 il rapporto tra età mentale ed età cronologica.

È chiaro che il Q. I. inteso in questo modo altro non è che un indice dello sviluppo intellettuale. Un bambino di 7, anni con un Q. I. di 125, è un bambino che ha uno sviluppo intellettuale pari a quello di un bambino di poco più di 9 anni; non è detto però che a 9 anni quel bambino abbia ancora un Q. I. di 125, cioè uno sviluppo intellettuale pari a quello di un bambino di quasi 11 anni.

Dopo l'apparizione del test di Binet si ha, soprattutto negli Stati Uniti, un'enorme fioritura di dibattiti accademici e di test di tutti i tipi. Nel 1919 J. B. Watson aveva chiaramente enunciato i principi del comportamento secondo il quale «obiettivo teorico di fondo della psicologia è la previsione e il controllo del comportamento». Anche il pensiero, considerato semplicemente come un linguaggio interno è comportamento. Ben pochi psicologi però concordavano sulla natura delle funzioni umane che i test tentavano di misurare. Nel 1921 una nota editoriale del «Journal of Educational Psychology» riferiva che era stato chiesto ai maggiori psicologi di dire che cosa intendevano per intelligenza. 17 di questi tra cui Terman, Thorndike, Thurstone, Yerkes risposero, ma non vi furono due risposte conciliabili tra loro. Non so se fu in quella occasione che E. J. Thorndike definì l'intelligenza come la qualità per la quale un gruppo di geni come Newton, Einstein, Leonardo da Vinci, Shakespeare e Rabelais differisce da altrettanti ospiti di un istituto per oligofrenici. Piercy, 40 anni dopo, dirà che coloro che più si occupano del problema dell'intelligenza si trovano stranamente imbarazzati dalla richiesta di una definizione e preferiscono scrivere un trattato.

Il medico tedesco Franz Joseph Gall, che insegnò a Parigi a partire dal 1807, divenne celebre per le sue ricerche di anatomia e fisiologia cerebrale che lo condussero alla formulazione di una dottrina fisiognomica nota come frenologia. Gall suddivise il cervello in regioni dette organi che presiedevano al controllo di una funzione psichica; nel caso della memoria verbale le localizzazioni, qui indicate con i numeri 33, erano le facce orbitali dei lobi frontali. Poiché a ogni organo cerebrale corrispondeva una bozza cranica Gall e il suo allievo J. C. Spurzheim pensavano che le qualità intellettuali di un individuo fossero deducibili dall'aspetto delle sue bozze. Questa teoria, ormai del tutto sorpassata, ebbe tuttavia il merito di formulare delle ipotesi sui rapporti tra funzioni corticali superiori e determinate aree corticali, rapporti che sono alla base della neuropsicologia.

Le analisi fattoriali

Da questi dibattiti sono nate due grandi teorie dell'intelligenza a tutt'oggi considerate valide; una in Inghilterra e l'altra in America.

In Inghilterra, il filosofo Herbert Spencer, ricollegandosi alla teoria dell'evoluzione di Darwin, propose una gerarchia evolutiva dei meccanismi mentali dove le funzioni più recenti e più complesse si costruivano fondandosi su quelle più vecchie e più semplici. Questa teoria da un lato ebbe molta influenza sul neurologo J. H. Jackson, che riconobbe una simile gerarchia di organizzazione all'interno del sistema nervoso, e dall'altro sullo psicologo C. Spearman e sulla sua concezione dell'intelligenza che espose nel 1927. Spearman notò che tutti i test di abilità cognitive, per quanto diversi, erano in correlazione positiva tra di loro; ciò gli suggerì l'esistenza di una qualche abilità generale che entrava in misura diversa nei vari test e che chiamò «g». Poiché nessuna delle correlazioni si avvicinava all'unità, egli pensò che per risolvere un test occorre un'abilità specifica (diversa da test a test) oltre all'abilità generale o «g». Spearman non aveva mai esplicitamente identificato «g» con l'intelligenza, ma tacitamente considerava «g» come l'abilità generale e questa convinzione gli veniva dall'osservazione che i test che maggiormente coinvolgevano «g» sono quelli che fanno maggiormente appello a un «alto livello di attività intellettuale». Quella di Spearman è una teoria gerarchica dell'intelligenza che pone il fattore «g» al di sopra delle altre abilità specifiche.

In America è stato invece favorito un altro tipo di analisi fattoriale che permetteva di evidenziare più fattori tra loro equivalenti. L. L. Thurstone applicò dapprima questo tipo di analisi allo studio delle attitudini e della personalità dove è ragionevole aspettarsi un certo numero di fattori pressappoco di uguale importanza. Sette anni dopo, nel 1945, sottopose 240 studenti a 56 test e isolò alcuni fattori che identificò con le abilità mentali primarie, tornando in un certo senso alle «facoltà» proposte dalla scuola frenologica, cioè dalla dottrina fisiognomica o teoria localizzatoria cerebrale fondata dal tedesco F. J. Gall nei primi anni del XIX secolo. Non vi è comunque nessuna garanzia che i fattori estratti con un'analisi matematica abbiano una loro realtà biologica (psicologica o neurologica). I fattori vengono però di solito intuitivamente interpretati sulla base del test che è maggiormente saturo di quel fattore; se, per esempio, test che appaiono spaziali sono altamente saturati con un dato fattore, questo verrà considerato un fattore spaziale, in pratica cioè si fa un salto logico: si estrae con un'analisi matematica un fattore e gli si attribuisce poi una realtà psicologica ritenendo che questi fattori riflettano alcune abilità di base.

Le basi neurologiche

Per alcune di queste teorie dell'intelligenza si può trovare un correlato neurologico; Gall e Spurzheim avendo constatato il particolare sviluppo dei lobi frontali negli uomini rispetto agli animali, ne dedussero che nei lobi frontali erano localizzate le funzioni più alte tipiche dell'uomo e, in particolare, l'intelligenza. Le lobotomie frontali o i vasti tumori frontali che distruggono gran parte dei lobi frontali non provocano però dei grossi deficit intellettuali e non confermano quindi l'ipotesi di Gall e Spurzheim.

Negli anni venti K. S. Lashley si dedicò allo studio dell'apprendimento nei ratti ai quali aveva asportato parti più o meno ampie di tessuto cerebrale, bilate-

FONTI DI VARIABILITÀ	GRADI DI LIBERTÀ	CHI2	P
MIGLIORATI × TEMPO	2	15,056	0,001
MIGLIORATI × TIPO	1	0,051	n. s.
MIGLIORATI × GRAVITÀ	1	16,344	0,001
MIGLIORATI × RIEDUCAZIONE	1	15,714	0,001

In questa tabella sono riportati i risultati dell'analisi della covarianza eseguita sui dati esposti in questa pagina in alto. L'analisi conferma quanto già detto: i pazienti rieducati migliorano significativamente di più di quelli non rieducati, mentre la gravità dell'afasia e la distanza dall'ictus hanno un effetto negativo sul miglioramento. Il tipo di afasia infine (fluente/non fluente) non ha un effetto significativo sul recupero.

Il test delle Matrici progressive di J. C. Raven consiste nel sottoporre al soggetto una serie di figure che devono essere completate con un tassello adatto scelto tra quelli rappresentati nella parte inferiore della pagina. La progressività del test dipende dal fatto che le figure da completare sono di difficoltà crescente. Poiché il test non richiede la mediazione del linguaggio anche afasici gravi possono ottenere un buon punteggio, purché abbiano mantenuto la capacità di fare deduzioni logiche.

ralmente; i ratti imparavano a muoversi in un labirinto con tanta maggior difficoltà quanto più estesa era la lesione indipendentemente dalla sede di asportazione purché fosse bilaterale e simmetrica. Lashley considerava apprendimento e intelligenza come equivalenti e riteneva inoltre che si poteva passare per analogia dal funzionamento del cervello del ratto a quello dell'uomo essendovi solo una differenza quantitativa; la sua conclusione era che il tessuto cerebrale è equipotenziale e che il deficit conseguente alla lesione dipende esclusivamente dalla quantità di tessuto leso («effetto massa»).

La teoria di Spearman dell'esistenza di un fattore intellettuale generale - che è una funzione del cervello intero e che si manifesta in tutte le prestazioni cognitive - è stata molto rafforzata da queste affermazioni di Lashley. Questi era un sostenitore del concetto di intelligenza come di una funzione del cervello in toto e nel 1949 disse che i test di intelligenza, come il test di Binet-Simon, basati su diversi livelli di età, identificano stadi di sviluppo - in apparenza qualitativamente diversi - che emergono durante la maturazione. Questi cambiamenti, secondo Lashley, devono però essere ricondotti a variazioni quantitative e non qualitative del sistema nervoso centrale e avanza l'ipotesi che nel cerebroleso si perdano aspetti qualitativamente diversi di una data abilità a seconda della quantità di cervello - necessaria a quell'abilità - rimasta intatta. Lashley riporta quindi tutte le differenze di comportamento che si ritrovano nella scala filogenetica e ontogenetica - oltre che nel cerebroleso - a differenze quantitative nello sviluppo del cervello e dichiara che l'unico fattore in correlazione con la complessità del comportamento è l'aumento di peso del cervello. La teoria di Lashley dell'equipotenzialità delle aree cerebrali non è però più sostenibile; da allora, infatti, vi sono stati importanti studi sperimentali e osservazioni cliniche che richiedono un radicale cambiamento nelle nozioni sulle basi neurologiche dell'intelligenza. Citerò soltanto due fatti: in nessuno studio sperimentale si è trovata una correlazione tra l'entità della lesione focale e la gravità del danno psicologico: in casi di emisferectomia infantile chirurgica, il livello di sviluppo intellettuale raggiunto può, in alcuni casi, essere appena al di sotto del livello normale, come è stato dimostrato da McFie nel 1961.

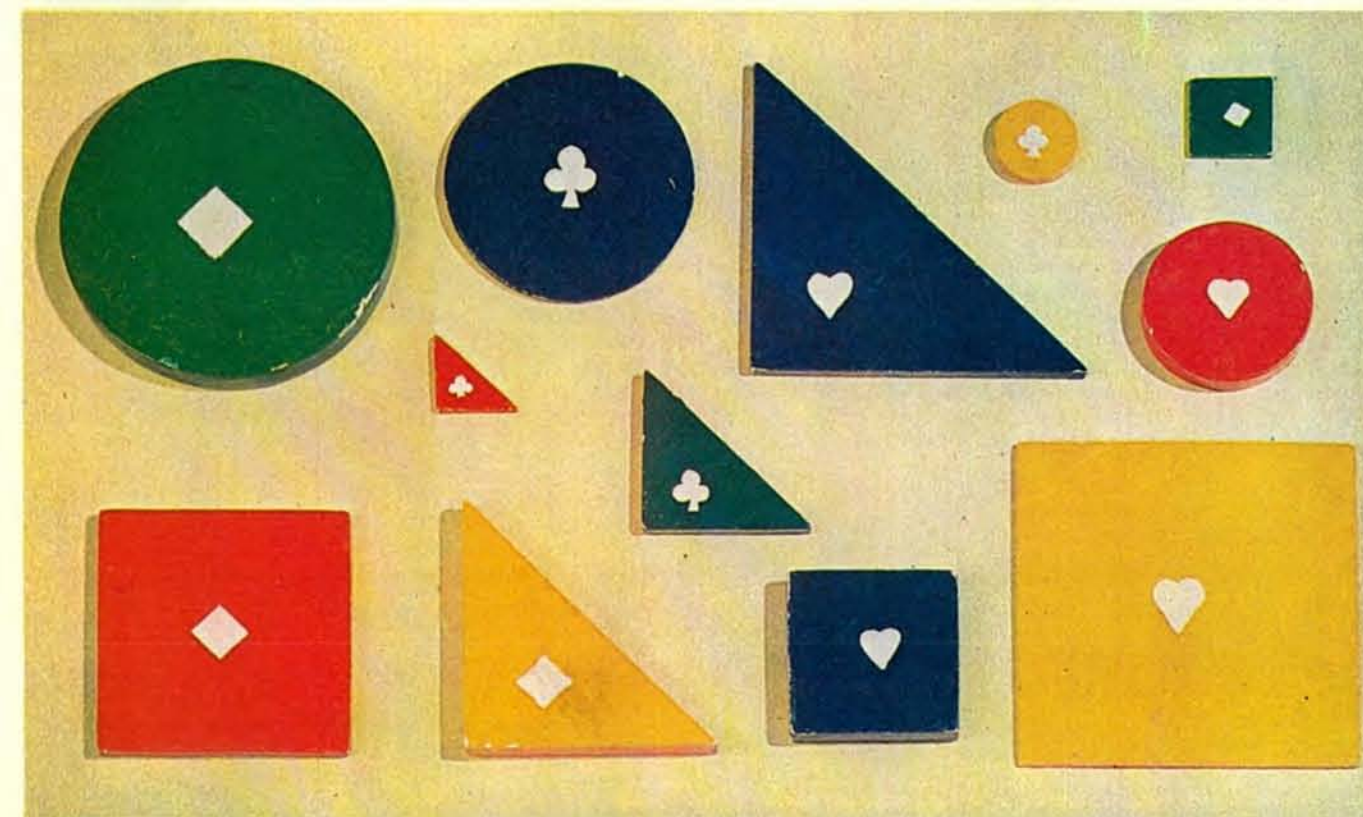
La teoria delle abilità primarie di Thurstone basata sull'analisi fattoriale trova il suo riscontro anatomico nel fatto che certe lesioni cerebrali focali provocano deficit intellettuali selettivi come le afasie, le agnosie (incapacità limitata a una sola modalità sensoriale di identificare figure od oggetti in assenza di deficit sensitivi primari), le aprassie (disturbo della gestualità volontaria in assenza di deficit motori) ecc., che possono, fino a un certo punto, essere appunto considerate equivalenti alle abilità primarie identificate da Thurstone. Vedremo, in seguito, che altri autori hanno creduto di riconoscere, in aree specifiche della corteccia cerebrale, zone particolarmente importanti per l'intelligenza.

Gli studi sperimentali

Come si può vedere da quanto è stato fin qui detto sull'intelligenza, negli stessi anni in cui i neurologi si chiedevano se l'afasia fosse un disturbo intellettuale, gli psicologi dibattevano che cosa fosse l'intelligenza. Il problema dei rapporti tra afasia e intelligenza era stato posto in maniera infelice e, nel corso degli ultimi 20 anni, si è tentato di affrontare diversamente il problema e di rispondere sulla base di dati sperimentali e non speculativi. I primi a usare test standardizzati per valutare l'eventuale presenza e gravità di un deficit intellettuale negli afasici sono stati T. H. Weisenburg e K. E. McBride nel 1935. Constatarono che gli afasici ottenevano risultati peggiori in tutte le prove verbali, ma, sottoposti a prove non verbali, i risultati ottenuti erano molto più variabili e in alcuni casi erano quasi normali. Zangwill riportando nel 1964 i risultati ottenuti da alcuni afasici ai test delle matrici progressive messi a punto da J. C. Raven concludeva che la capacità di risolvere un test di intelligenza, che sembra coinvolgere un ragionamento formale normalmente mediato dal linguaggio, è relativamente poco compromessa in qualunque forma di afasia. Contemporaneamente Alajouanine e Lhermitte portano dei dati sui risultati al test di Wechsler non verbale di un gruppo di 218 afasici; solo nel 25 per cento dei casi il quoziente intellettuale per le prestazioni manuali era inferiore a 80, ma la percentuale negli aprassici costruttivi - dove per aprassia costruttiva si intende l'incapacità di copiare disegni geometrici, astratti o figurati, legata a un deficit dell'esplorazione, trasposizione o ricostruzione - era dell'80 per cento benché nell'intero gruppo gli aprassici costruttivi rappresentassero solo il 15 per cento. Sembrerebbe quindi che il cattivo rendimento alla scala manuale del Wechsler-Bellevue sia dovuto alla presenza di aprassia costruttiva e non all'afasia di per sé. Risultati analoghi hanno ottenuto Arrigoni e De Renzi che hanno trovato che gli aprassici costruttivi con lesioni dell'emisfero sinistro ottengono risultati peggiori alle Matrici progressive di Raven di tutti gli altri gruppi. Altri autori (Orgass, Hartje, Kerschensteiner e Poeck), partendo dal presupposto che un danno cerebrale porta di per se stesso a un deficit intellettuale e che gli afasici hanno per definizione un danno cerebrale, hanno confrontato un gruppo di cerebrolesi afasici con altri cerebrolesi sinistri senza afasia e cerebrolesi destri. I sinistri non afasici ottennero alle prove del Wechsler-Bellevue punteggi superiori ai cerebrolesi destri che, a loro volta, ottennero punteggi significativamente superiori agli afasici. I profili dei punteggi ottenuti nelle cinque sottoprove sono leggermente diversi nei vari gruppi.

Il quoziente di deviazione

Nella maggior parte dei lavori sperimentali sull'intelligenza degli afasici, è stato usato il test di Wechsler-Bellevue (o il Wechsler Adult Intelligence Scale - WAIS -che ne è una versione successiva); il risultato ottenuto alle varie prove viene espresso con un numero che è detto quoziente intellettuale. È importante notare che anche se si usa lo stesso termine per il risultato ottenuto al test di Binet e al test di Wechsler (o a test analoghi per adulti), i concetti sottostanti sono estremamente diversi. Abbiamo detto che nel primo caso Q. I. non è altro che un indice dello sviluppo intellettuale; questo sviluppo però non prosegue tutta la vita, ma si arresta, a seconda degli autori, tra i 15 e i 25 anni. Dopo quest'età non ha dunque alcun senso calcolare il Q. I. in base all'età mentale; usando questo sistema si arriverebbe all'assurdo che una persona di 30 anni, per esempio, di intelligenza media, avrebbe un'età mentale inferiore alla sua età cronologica e ciò perché vi è un arresto della crescita mentale, e l'età mentale massima per ogni prova sarebbe quella in cui la prova è risolta dal 100 per cento dei soggetti. Nel 1958 D. Wechsler ha abbandonato il concetto di età mentale nei test per adulti e ha introdotto il concetto di un Q. I. che indica soltanto la collocazione di una persona in rapporto ai suoi coetanei. Wechsler ha calcolato, per ogni livello di età, i risultati (in punteggi grezzi) a un test di intelligenza (il Wechsler-Bellevue dapprima, 1937, e il WAIS in seguito, 1955); ne ha ottenuto delle curve normali e ha attribuito alla media il valore arbitrario di 100; ha inoltre fissato il valore della deviazione standard (DS) in 15. I Q. I. calcolati in questo modo sono anche noti come quozienti di deviazione; tra i valori 85 e 115 (media ± 1 DS) sappiamo, in base alla distribuzione normale, che è compreso circa il 68 per cento della popolazione e il 95 per cento tra 70 e 130 (media ± 2 DS).

Dopo questa digressione sul significato del Q. I. negli adulti, riportiamo ancora i risultati di alcuni lavori sperimentali. Alcuni ricercatori italiani, E. De Renzi, P. Faglioni, M. Savoiardo e L. A. Vignolo sottoposero nel 1966 un gruppo di controllo, un gruppo di cerebrolesi destri, cerebrolesi sinistri non afasici e cerebrolesi sinistri afasici a un test di pensiero categoriale (una versione modificata dal test di Weigl illustrata qui sotto che permette raggruppamenti in base a 5 diversi criteri) e trovarono che sono i cerebrolesi sinistri afasici i soggetti che avevano un rendimento peggiore e che i punteggi ottenuti al test di Weigl erano correlati con quelli di un test di comprensione. In un lavoro del 1973, l'autrice e alcuni collaboratori trovarono che il rendimento degli afasici alle Matrici progressive di Raven 47 era significativamente inferiore a quello di un gruppo di controllo. Tuttavia, la correlazione tra il punteggio alle Matrici progressive e il punteggio di espressione e comprensione orale era praticamente zero e non si poteva quindi spiegare il cattivo rendimento degli afasici come conseguente all'afasia. L'ipotesi degli autori era che, per quanto riguarda l'emisfero sinistro, vi fosse una sovrapposizione anatomica tra le aree del linguaggio e le aree deputate all'«intelligenza».

Il test di Weigl è stato messo a punto per mettere alla prova la capacità di astrazione dei pazienti afasici. Questa versione del test è stata modificata in modo da portare da tre a cinque il numero di compiti eseguibili. All'esaminato viene infatti richiesto di raggruppare i pezzi riprodotti in figura in base a una delle loro caratteristiche come la forma, il colore, il seme delle carte da gioco, la dimensione e lo spessore.

Sempre per fornire una valutazione della capacità di riconoscere le categorie degli oggetti, al paziente viene chiesto di accoppiare una figura presentata con un oggetto messo a sua disposizione. Nell'esempio qui illustrato per l'accoppiamento sono state utilizzate due chiavi di foggia assai diversa.

I francesi Messerli e Tissot studiarono l'intelligenza negli afasici partendo da presupposti diversi: essi ritenevano che così come nell'afasia si ha una disintegrazione del linguaggio verso forme più infantili, l'intelligenza degli afasici può anch'essa regredire a stadi di sviluppo precedenti. Rifacendosi alle teorie di Jean Piaget, hanno sottoposto un gruppo di afasici a prove corrispondenti a stadi di sviluppo diversi concludendo che l'afasia non colpisce l'intelligenza per sé, ma disturba quelle operazioni intellettuali per le quali è indispensabile l'influenza mediatrice del linguaggio.

Gli strumenti

E qui sorge un ulteriore problema: l'intelligenza, qualunque cosa essa sia, non è mai direttamente conoscibile; ciò che si può vedere e valutare è solo il comportamento di una persona in una data situazione. È chiaro che questo comportamento per estrinsecarsi deve utilizzare degli strumenti che possono essere di varia natura; il modo più usato per valutare l'intelligenza di una persona è quello di sottoporle dei problemi verbali di varia natura; per rispondere occorre che questa persona possa usare il linguaggio. Benché test apertamente verbali non siano più usati per valutare l'intelligenza degli afasici, il problema non è ovviamente così semplice. Il cattivo rendimento degli aprassici costruttivi alla scala manuale del test di Wechsler-Bellevue è determinato dal fatto che il test di Wechsler è, almeno in parte, un test di prassia o è effettivamente dovuto a un deficit intellettuale? Le Matrici progressive sono un test di abilità generale o in esse vi è, come riteneva Raven, anche un fattore spaziale? E allora chi cade a questo test, cade perché poco intelligente o perché poco «spaziale»? Il problema quindi è quello di riuscire a distinguere tra lo strumento che l'intelligenza utilizza per manifestarsi e l'intelligenza stessa. È un problema che è praticamente impossibile risolvere ed è tanto più difficile se lo strumento usato dall'intelligenza è il linguaggio. Nel bambino piccolo, intelligenza e linguaggio hanno uno sviluppo che è, almeno in parte, separato. Piaget vede i vari stadi dello sviluppo come costruiti sulle fondamenta formate dai precedenti e più semplici livelli di sviluppo; il linguaggio si sviluppa solo sulla base dell'intelligenza senso-motoria che è quindi, almeno inizialmente, totalmente separata dallo sviluppo del linguaggio. In un secondo momento, però, è l'uso del linguaggio che permette di giungere allo stadio delle operazioni logico-formali. Vedremo più avanti come gli ultimi lavori sulla comprensione degli afasici tornano a questo problema del confine tra linguaggio e concetti (cioè tra linguaggio e quell'operazione intellettuale di astrazione che permette di passare dall'oggetto contingente «mela» a tutte le mele esistenti, future o solo possibili).

FONTI DI VARIABILITÀ	PUNTEGGI		
	GRADI DI LIBERTÀ	F	F
PRESENZA DI AFASIA	2/161	13,513	0,0001
TIPO DI AFASIA (FLUENTE O NON FLUENTE)	2/161	1	n. s.
GRAVITÀ DELL'AFASIA	2/161	1	n. s.
TIPO PER GRAVITÀ	2/161	1,125	n. s.

In questa tavola sono riportati i risultati dell'analisi della covarianza dei punteggi ottenuti ai test di Raven e Wechsler corretti, oltre che per età, scolarità e durata di malattia, per volume lesionale e punteggio di aprassia (ideomotoria e costruttiva). La presenza di afasia risulta altamente significativa (gli afasici cioè hanno punteggi inferiori ai non afasici), ma il tipo di afasia, la sua gravità e la loro interazione non influiscono significativamente sui risultati ai test.

I concetti

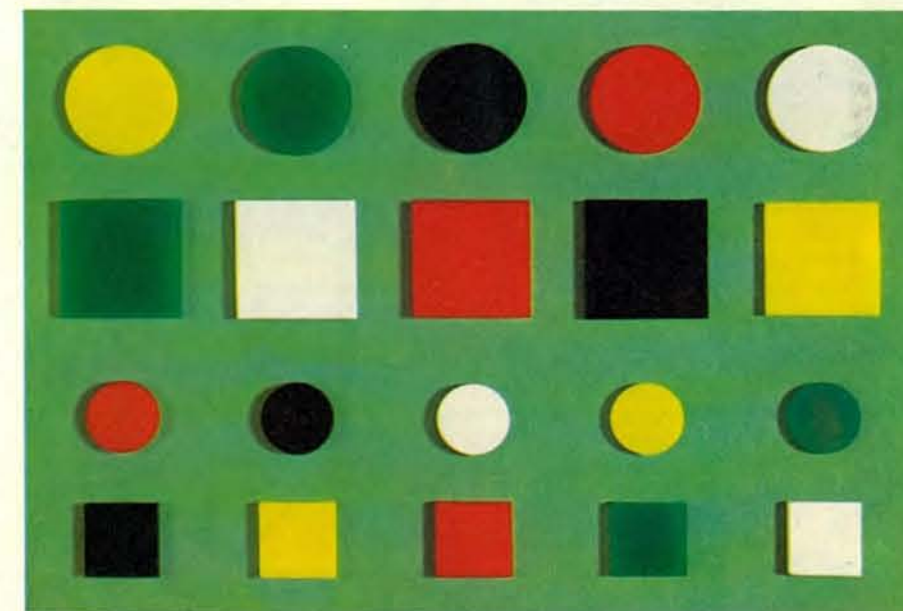
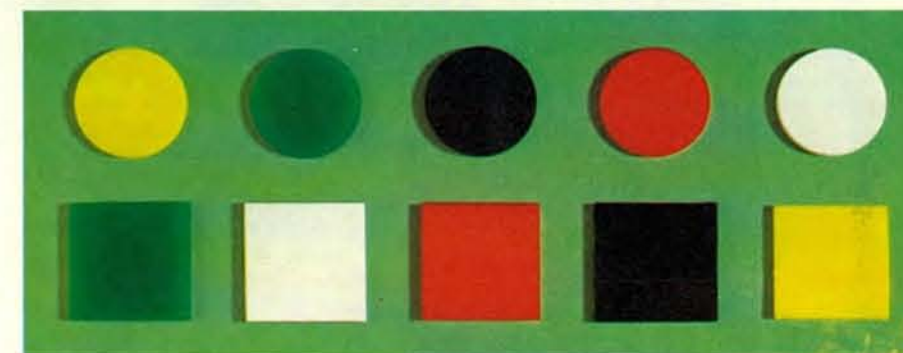
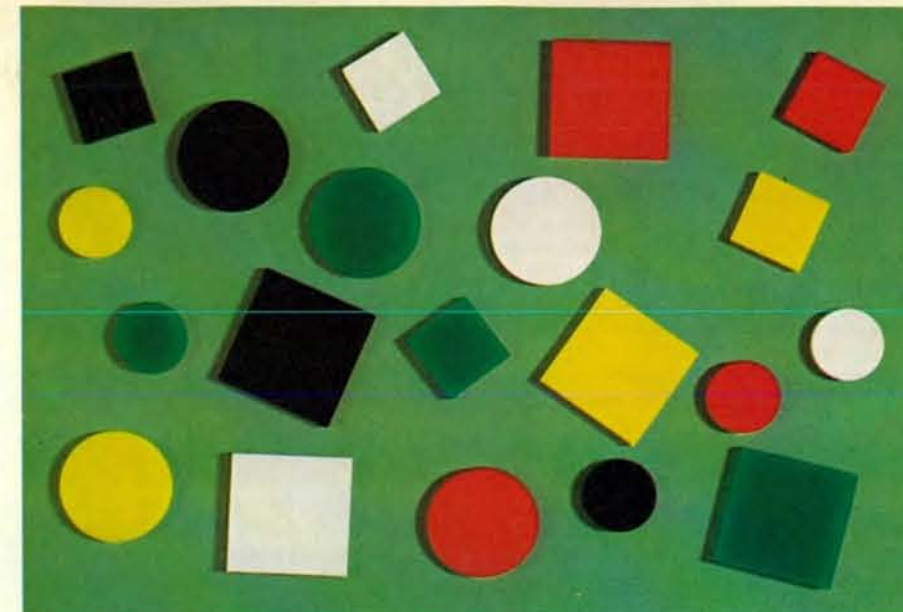
Lasciamo per il momento da parte il problema di separare l'intelligenza dal comportamento attraverso il quale si manifesta e vediamo che cosa ci possono dire sui deficit intellettuali degli afasici una serie di altri lavori il cui principale oggetto di ricerca era quasi sempre l'incapacità di riconoscere gli oggetti. In una prima ricerca di H. Spinnler e L. A. Vignolo, si voleva verificare la capacità di soggetti cerebrolesi destri e sinistri di associare un rumore significativo alla sua fonte sonora (per esempio un suono di campana alla figura di una campana); in un'altra ricerca di E. De Renzi, A. Pieczuro e L. A. Vignolo, la capacità di associare a un oggetto (martello) i gesti generalmente compiuti nel maneggiarlo (martellare). E. De Renzi e H. Spinnler osservarono invece la capacità di soggetti cerebrolesi ad accoppiare un disegno in bianco e nero con il suo colore caratteristico (ciliege: color rosso). In tutti questi compiti i soggetti afasici si dimostrarono inferiori ai cerebrolesi destri e ai soggetti di controllo.

Qual è il denominatore comune a queste varie prove? Sembra che si possa riconoscere che tutte, per essere ben risolte, richiedono una medesima operazione mentale: passare dall'oggetto presentato al concetto di quell'oggetto che comprende, per esempio, sia il gesto di martellare che il martello, sia il colore rosso che il disegno della ciliegia. Sulla base di questi risultati sperimentali si può quindi forse affermare che il concetto sottostante all'oggetto presentato è, negli afasici, impoverito o comunque distorto.

Una riprova di questo fatto l'abbiamo in una recente ricerca eseguite nel 1978 da P. Whithouse, A. Caramazza e E. Zurif; Questi autori presentano a soggetti di controllo e cerebrolesi sinistri una serie di disegni: una tazza, un bicchiere e delle deformazioni successive del primo disegno - tazza - che viene a rassomigliare sempre più al secondo disegno - bicchiere -. Compito del soggetto è dire se il disegno presentato è una tazza o un bicchiere. Questa serie di disegni viene presentata 3 volte: senza nulla, con acqua o con caffè. Secondo gli psicologi un concetto si basa sulle caratteristiche di forma (percettive) e d'uso (funzionali) dell'oggetto (un bicchiere è un oggetto fatto così e così che serve per bere). Sulla base delle risposte date dai soggetti di controllo gli autori hanno costruito dei profili di risposta e hanno trovato che le risposte sono coerenti e tengono conto oltre che dei dati percettivi anche della funzione (viene considerato tazza un dato disegno se contiene caffè, ma bicchiere se contiene acqua). Le risposte date dagli afasici non fluenti sono sovrapponibili a quelle date dai controlli, mentre quelle date dagli afasici fluenti differiscono: sono meno coerenti e non tengono conto del contesto. Gli autori concludono che gli afasici fluenti possiedono concetti sottostanti alle parole più poveri e vaghi dei soggetti di controllo.

Come si vede da queste ultime ricerche, il problema dei rapporti tra intelligenza e afasia è andato mutandosi nel corso degli anni. All'inizio si riteneva che un deficit intellettuale globale fosse all'origine del disturbo afasico; negli stessi anni gli psicologi andavano maturando varie teorie dell'intelligenza e mettevano a punto vari test. Utilizzando i test di intelligenza si è poi voluto verificare se, in qual misura e quali afasici presentassero dei deficit intellettuali. In questi ultimi anni ci si è piuttosto chiesti se ci sono degli specifici deficit intellettuali negli afasici, e l'aspetto più studiato è stato quello dei concetti sottostanti le parole. Il concetto è una operazione mentale di categorizzazione e sembra che in qualche modo l'afasico, perlomeno gli afasici fluenti, non abbiano più a disposizione dei concetti chiari e sufficientemente ricchi. Vi è ormai accordo generale tra tutti coloro che si occupano dell'afasia sul fatto che l'afasico non è semplicemente una persona sana che tenti di comunicare in una lingua che conosce imperfettamente, ma c'è qualcosa in più che non si riesce ancora a identificare. In un recente lavoro condotto in questi ultimi anni e in corso di stampa, abbiamo tentato di rispondere almeno in parte a queste domande.

I risultati dei test

173 cerebrolesi sinistri sono stati sottoposti a un esame del linguaggio, al test dei gettoni, ai test per la valutazione della prassia ideomotoria e della prassia costruttiva, alla scala manuale del test di Wechsler-Bellevue e alle Matrici progressive di Raven. Per ognuno di questi pazienti era inoltre disponibile una misura del volume della lesione calcolato sulla base della tomografia assiale computerizzata. In una prima analisi della covarianza sono stati confrontati i risultati dei cinque gruppi sperimentali ottenuti dividendo i soggetti in base alla presenza, tipo e gravità di afasia (cerebrolesi sinistri non afasici, afasici non fluenti gravi, afasici non fluenti di media gravità; afasici fluenti gravi ed afasici fluenti di media gravità); la presenza e la gravità di afasia (non il tipo) si sono dimostrate significative nei confronti dei risultati al Raven e al Wechsler. Quando, però, con una seconda analisi della covarianza, si sono introdotte come variabili correggenti il volume lesionale e i punteggi di aprassia è risultato che

Il test dei gettoni è stato messo a punto presso la Clinica neurologica dell'Università di Milano per valutare l'entità dei disturbi afasici della comprensione orale. Anche questo test è influenzato dal livello culturale del soggetto esaminato. Esso dà la possibilità di eseguire parecchie prove di difficoltà crescente. Dapprima l'esaminatore dispone i gettoni in ordine sparso come appaiono nella figura in alto e al soggetto viene richiesto di indicare per esempio «un quadrato rosso» o un «cerchio verde». L'esaminatore passa poi a disporre solo i gettoni grandi come appare nella figura al centro e chiede per esempio al soggetto di indicare «il quadrato giallo e il cerchio verde» aumentando così la complessità dell'ordine. Nelle prove finali tutti i gettoni vengono disposti come appare nella figura in basso. All'esaminato si può chiedere di indicare «il cerchio bianco piccolo e il quadrato rosso grande» o compiti grammaticalmente più complessi come «porre un cerchio giallo sopra un quadrato rosso» oppure «toccare un quadrato bianco quando l'esaminatore tocca un cerchio rosso». In quest'ultima prova anche i soggetti normali compiono errori.

l'effetto della gravità di afasia è scomparso, mentre restava altamente significativo l'effetto della presenza di afasia. Considerando il significato della correzione apportata dall'aprassia costruttiva, si vede che questa è altamente significativa; gli aprassici costruttivi ottengono risultati peggiori sia al Raven, sia al Wechsler, anche quando si è controllato l'effetto dell'afasia e del volume della lesione. La correlazione con il volume di lesione si è sempre dimostrata piuttosto bassa. In conclusione, questi dati indicano che non è il volume della lesione responsabile del basso rendimento al Raven e al Wechsler, come all'interno degli afasici non lo è la gravità di afasia (del resto anche nei lavori precedenti con il test dei gettoni, la correlazione era sempre risultata piuttosto bassa indicando così che non era l'afasia di per sé a causare il cattivo rendimento degli afasici nei vari test d'intelligenza di volta in volta usati). Sono invece gli aprassici costruttivi i soggetti che hanno il peggior rendimento ed è difficile spiegare il loro cattivo rendimento con il fatto che Raven e Wechsler impegnano la prassia come strumento perché sono test molto eterogenei tra loro e certamente non tutte le prove del Wechsler richiedono la stessa prassia, mentre i risultati sono gli stessi.

La teoria di Hebb

Partendo dal problema dei rapporti tra eredità e ambiente da un lato e l'osservazione che le lesioni cerebrali hanno conseguenze diverse nei bambini e negli adulti, D. O. Hebb in alcuni lavori degli anni cinquanta ha proposto una nuova teoria dell'intelligenza, psicologica e neurologica. La letteratura psicologica di quegli anni sottintendeva che «intelligenza» ed «efficienza del cervello» fossero equivalenti psicologicamente e quanto a substrato neuro-anatomico. Il quoziente intellettuale era considerato una misura della potenza del cervello, a sua volta considerata immutabile fino alla vecchiaia. Secondo Hebb, l'esperienza ha un ruolo importante nello sviluppare le attività cognitive e un test di intelligenza valuta essenzialmente queste attività e non l'adeguatezza fisiologica del cervello. Hebb viene così a distinguere una intelligenza A e una intelligenza B; la prima si riferisce al potenziale innato di sviluppo della capacità intellettuale di formare concetti, mentre la seconda si riferisce al livello raggiunto a uno stadio successivo di sviluppo. L'intelligenza A, da un punto di vista fisiologico, può essere equiparata all'efficienza cerebrale e determinare la velocità di sviluppo e il livello intellettuale di un bambino in un ambiente ottimale; il livello intellettuale effettivamente raggiunto, o intelligenza B, dipenderà dalla opportunità di formare concetti che l'ambiente offre. L'effetto dell'ambiente sembra essere quello di stabilire delle attività concettuali che si sviluppano poi gerarchicamente l'una sull'altra: solo chi possiede i concetti *1* e *2* è successivamente in grado di sviluppare il concetto *3*, non prima. Questo sviluppo di processi concettuali, secondo Hebb, continua negli anni; anche quando, con l'avanzare dell'età, il cervello perde neuroni, il suo livello di funzionamento può ancora aumentare; i concetti nuovi, infatti, nascono dai concetti vecchi a valanga e l'impiego di neuroni necessario a mantenere questo meccanismo è minimo.

Hebb tenta anche di dare una interpretazione neurologica del funzionamento dell'intelligenza; parte dall'osservazione che un danno cerebrale in un bambino tende a provocare un deficit intellettuale più diffuso di quanto non faccia lo stesso danno nell'adulto; negli adulti i deficit conseguenti a una lesione cerebrale sono più specifici. Secondo Hebb, quindi, per sviluppare una data abilità intellettuale è necessaria una maggior quantità di tessuto cerebrale di quanta non ne occorra per «sostenerla» una volta che questa sia acquisita; in un adulto una vasta lesione cerebrale non riduce necessariamente l'intelligenza in proporzione alla sua entità, perché prima della lesione si è avuto uno sviluppo concettuale. L'intelligenza B, una volta sviluppata, richiede cioè una piccola parte di tessuto cerebrale e non viene grandemente danneggiata da una lesione; all'opposto, l'intelligenza A, cioè il potenziale intellettuale, la capacità di acquisire nuove capacità intellettuali, è gravemente danneggiata da una lesione cerebrale.

Il comportamento degli afasici

Prima di tentare di interpretare quanto detto fin qui, vorrei aggiungere un'ultima osservazione. La lunga pratica quotidiana con gli afasici, l'osservazione del loro comportamento in varie situazioni della vita, i resoconti dei familiari, così come i rapporti ottenuti dai colleghi di lavoro per quei pazienti che sono stati in grado di riprendere un'attività lavorativa, tutto porta ad affermare che il comportamento degli afasici, tenuto conto delle loro difficoltà linguistiche, è quasi sempre perfettamente adeguato alla situazione, anche per quei pazienti che ottengono punteggi estremamente bassi ai vari test di intelligenza, anche non verbali. In ciò, il comportamento di questi pazienti è nettamente diverso da quello dei dementi. Il concetto di demenza neurologica non è semplicemente il contrario di quello di intelligenza, ma è quello di una malattia neurologica acquisita, il cui prototipo è la malattia di Alzheimer nelle sue varianti giovanile, presenile e senile. La demenza può essere considerata come un quadro clinico evolutivo, maligno, di insufficienza d'organo (il cervello) con progressiva risoluzione di ogni possibile contatto sociale e, da ultimo, della stessa sopravvivenza. Dire da che punto in poi della malattia una persona possa essere etichettata come demente non è frutto della prestazione ai test, ma è essenzialmente una questione sociale centrata sulla capacità di mantenersi autonomo: il paziente con insufficienza cerebrale progressiva cesserà di essere autosufficiente molto prima in una società industriale che in una pastorale che richiede un contributo raziocinante al proprio comportamento sociale notevolmente minore. Comunque prima o poi il comportamento del paziente sarà caratterizzato da sempre più frequenti fallimenti neuro-psicologici (dimenticherà, si perderà, non saprà più usare strumenti di tutti i giorni ecc.) che di fatto lo escluderanno nel giro di 1-3 anni dal contesto sociale in cui era inserito.

Tornando all'interpretazione dell'intelligenza di Hebb possiamo forse concludere quanto esposto fin qui, dicendo che negli afasici che presentano lesioni focali, il livello intellettuale acquisito o intelligenza *B*, che nell'adulto è sottesa da un tessuto cerebrale relativamente piccolo, è conservata e ciò spiega il fatto che il loro comportamento è solitamente perfettamente adeguato alla situazione e che sono tuttora in grado di svolgere compiti anche complessi, purché siano compiti che già svolgevano. La loro caduta ai test è invece spiegata con una perdita di potenziale intellettuale, di intelligenza A, quella cioè che permette di acquisire nuove capacità intellettuali e che è, secondo Hebb, molto sensibile al danno cerebrale. Nei test, infatti, si richiede al soggetto di risolvere compiti nuovi e inusuali che sembrerebbero, quindi, richiedere il funzionamento dell'intelligenza A.

Questa spiegazione sembrerebbe confermata anche dal fatto che se l'afasia è insorta in giovane età (dopo l'acquisizione del linguaggio, ma ancora in età scolare), mentre il recupero del linguaggio è di solito buono (migliore che negli adulti), questi ragazzi continuano ad avere difficoltà a scuola dove si richiede loro di acquisire nuovi concetti e relazioni basandosi sul loro potenziale intellettuale, cioè sull'intelligenza A. Abbiamo già detto come secondo Hebb un danno cerebrale in un bambino tenda a provocare un danno intellettuale molto più diffuso che nell'adulto.

Nei dementi infine, in cui la lesione non è focale, ma diffusa, ci sarebbe, oltre a una perdita dell'intelligenza A come nei cerebrolesi focali, anche una perdita dell'intelligenza B e ciò spiegherebbe l'inadeguato comportamento dei dementi anche in situazioni a loro familiari.

Da quanto detto fin qui, emerge chiaramente che non si sa con certezza che cosa si misuri, nei cerebrolesi, con test di intelligenza e quindi perché questi vengano regolarmente applicati.

Una possibile risposta è che se vogliamo, negli afasici, valutare tutte quelle funzioni che possono essere compromesse da una lesione emisferica sinistra, un test di intelligenza ha la stessa ragione d'essere di un test, per esempio, per la prassia ideomotoria. D'altro canto, una ragione più importante - anche se finora non confermata sperimentalmente - potrebbe essere che un buon rendimento a un test d'intelligenza può avere un valore prognostico positivo sul recupero del linguaggio.

I monumenti megalitici

Questi complessi di enormi pietre, il più noto dei quali è Stonehenge, si trovano a migliaia in Europa. La loro datazione è rimasta a lungo incerta, ma ora si hanno elementi per attribuirli al periodo neolitico

di Glyn Daniel

Tra i resti più suggestivi dell'antico paesaggio europeo vi sono i numerosi monumenti preistorici in pietra. Questi megaliti hanno per lungo tempo suscitato l'interesse e la curiosità del pubblico in generale e di antiquari e archeologi in particolare, inducendo questi ultimi a ricercare e a descrivere in modo corretto la natura, lo scopo, il contesto e l'età di queste strutture. Per citare solo due esempi, ricordiamo fra i più notevoli resti della Francia preistorica gli allineamenti di pietre della regione di Carnac, nella Bretagna meridionale, ove oltre 3000 menhir sono ancora in piedi, in linee parallele che si estendono per circa 6 chilometri e mezzo; e il Grand Menhir Brisé, nei pressi di Locmariaquer, ora spezzato, ma originariamente alto 22 metri.

Ma il più famoso tra tutti i monumenti megalitici è senza dubbio Stonehenge, che sorge nella pianura del Wiltshire, nell'Inghilterra meridionale. Visitato annualmente da migliaia di persone, come attrazione turistica è secondo soltanto alla Torre di Londra. Possiede una letteratura più ampia di qualsiasi altro sito archeologico nel mondo, comprese le piramidi d'Egitto e le grandi statue dell'isola di Pasqua, o i siti leggendari come Atlantide. Il numero di libri su Stonehenge e sugli altri monumenti megalitici dati alle stampe nell'ultimo decennio dà la misura del perdurare dell'interesse verso questi resti preistorici.

Purtroppo dà anche una dimostrazione assai chiara della fantasia, dell'illusione e della credulità di molti autori, e dell'abissale ignoranza di molti che si dicono archeologi, ma che, a essere severi, possono essere definiti soltanto seguaci dell'immaginazione. Non si tratta di un fenomeno nuovo. Nel lontano 1911 il libro di G. Elliot Smith *The Ancient Egyptians (Gli antichi egiziani)* ipotizzò la provenienza di tutti questi antichi monumenti europei dalle rive del Nilo. Simili esercizi di immaginazione continuano. Recentemente, nel 1977, Euan MacKie con *The Megalith Builders (I costruttori di megaliti)* affermò che essi erano opera di esperti dell'Egitto predinastico e di Sumer. Altri, tra i quali Erich von Dänichen, ritengono che i costruttori di megaliti fossero dei navigatori provenienti dallo spazio. Ora si è anche diffusa la credenza che questi monumenti siano stati costruiti con finalità di carattere astronomico, e circolano liberamente termini come «astroarcheologia» e «archeoastronomia». Gettiamo ora uno sguardo più prudente ed equilibrato su queste strutture, nel contesto delle nostre attuali conoscenze dell'Europa antica.

Lo studio dei megaliti

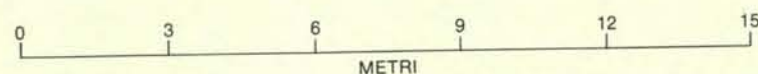
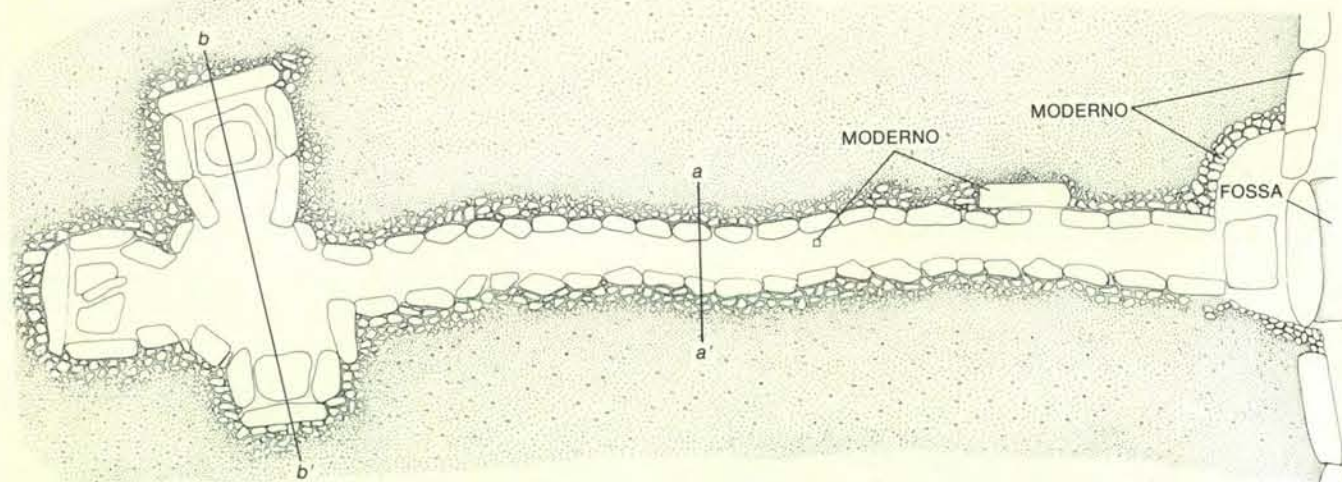
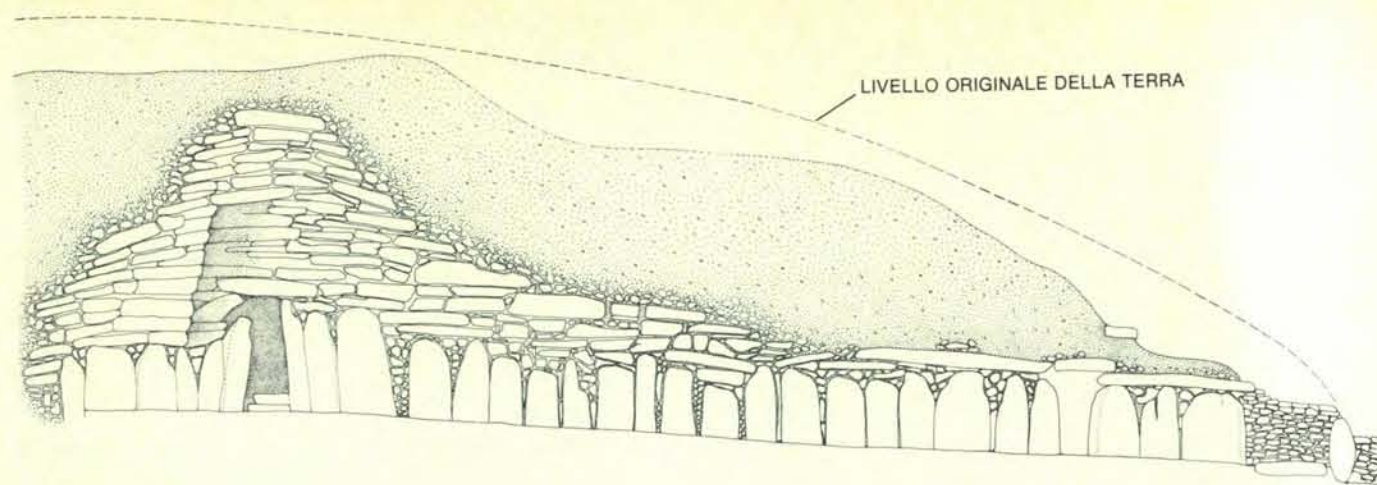
Nel 1849, in un libro intitolato *Cyclops Christianus,* un professore di Oxford, Algernon Herbert, coniò la parola megalito (dal greco *megas,* grande e *lithos,* pietra). La parola fece presa. Sebbene nel 1872 James Fergusson, uno studioso scozzese di architettura, avesse intitolato il suo libro *Rude Stone Monuments in All Countries (Monumenti in pietra grezza in tutti i paesi)* anch'egli parlò di megaliti. Altrettanto fece T. E. Peet nel suo *Rough Stone Monuments (Monumenti in pietra grezza),* pubblicato 40 anni più tardi.

Queste opere pionieristiche determinarono l'avvio di uno studio appropriato dei megaliti, ma introdussero anche elementi di fondamentale confusione: coloro che, come Fergusson, si dedicarono ai viaggi fuori dell'Europa, trovarono grandi monumenti in pietra in Algeria, in Palestina, in Etiopia e nel Sudan; inoltre nel Caucaso, in Persia, nel Belucistan, nel Cashmir e nell'India centrale e meridionale. (In India i monumenti megalitici del Deccan, molti dei quali con «aperture a portello», simili a quelle delle tombe megalitiche europee, furono il primo oggetto di interesse per Fergusson nell'intraprendere le sue ricerche comparative.) L'elenco non termina qui. Sono stati trovati megaliti nell'Assam, a Sumatra e in alcune isole del Pacifico come Malekula nelle Nuove Ebridi. (Le statue di pietra dell'isola di Pasqua, sebbene siano senza dubbio di grandi dimensioni, non sono megaliti secondo l'accezione generale del termine.) In Giappone tombe megalitiche vennero costruite a partire dal II secolo a.C. fino al VII secolo d.C., quando l'imperatore Kotoku ne proibì la costruzione considerandola uno spreco di energia. Anche le civiltà precolombiane del Nuovo Mondo praticarono questo tipo di architettura.

L'applicazione del termine megalito a tutti questi monumenti, rinvenuti nei paesi più diversi, appartenenti a differenti periodi e a contesti culturali differenti, diede luogo a una assurda convinzione: quella che tali strutture fossero geneticamente collegate, che fossero opera di una razza o di un popolo megalitico. Questa idea è stata ormai da tempo abbandonata, ed è stato ampiamente riconosciuto che l'analogia delle strutture megalitiche esistenti nelle diverse parti del mondo è dovuta unicamente all'analogia dei materiali, usati in modo simile. Questo parallelismo si rivela particolarmente sorprendente nelle camere o stanze megalitiche in cui sono incorporati gli elementi fondamentali di quella che è nota come architettura a trabeazione (dal latino *trabes,* trave). Questo tipo di struttura richiama alla mente la costruzione infantile di una casetta fatta di carte o di cubetti: lastre di pietra poste in piedi (ortostati) e altre poste trasversalmente sulle pietre verticali come elementi di copertura: Le possibili soluzioni architettoniche sono limitate, cosicché non c'è da sorprendersi se una camera megalitica, in Francia o in Irlanda, risalente al III millennio a.C. possa assomigliare a una camera megalitica dell'India meridionale appartenente alla fine del I millennio a.C.

Un altro errore abbastanza comune si deve alla confusione fra architettura megalitica e architettura ciclopica. Anche quest'ultima impiega pietre di grandi dimensioni; ma queste sono accuratamente accostate le une alle altre, anche se irregolari nella forma, e sono generalmente poste a strati. L'architettura ciclopica è reperibile sia nel Nuovo Mondo (ad esempio nelle strutture Inca del Perù) sia nel Vecchio Mondo (ad esempio nelle cittadelle greche di Micene e Tirinto o nei nuraghi sardi).

Il grande allineamento di Carnac, nella Bretagna meridionale, comprende tre distinti viali di menhir. In questa fotografia aerea è visibile in tutta l'estensione il gruppo di Kermario, formato da 10 file circa parallele di menhir, scavati nel granito, che si estendono per 1200 metri.

Il monumento più famoso della tradizione megalitica, Stonehenge, è un anello circondato da un argine e da un fossato, situato nella piana del Wiltshire nell'Inghilterra meridionale. L'anello fu eretto in fasi successive, a partire dal 2800 a.C. e fu completato poco dopo il 1100 a. C.

La camera sepolcrale più grande della tradizione megalitica è Newgrange, nella contea di Meath nell'Eire, qui illustrata in alzato *(in alto)* e in pianta e sezioni *(in basso)*. Lo stretto passaggio che conduce alla camera sepolcrale al centro del grande tumulo è lungo circa 18 metri; la volta aggettante della camera è alta 6 metri *(si veda la fotografia della pagina a fronte)*. Newgrange fu costruita intorno al 2500 a.C.

Le strutture megalitiche europee rientrano in quattro principali categorie. La prima è rappresentata dal menhir, o pietra ritta singola. Il termine deriva dal gallese *maen*, pietra e *hir*, lungo. La Bretagna è ricca di menhir, che variano in lunghezza da 1 a 6 metri. Una notevole eccezione è costituita dal Grand Menhir Brisé, una pietra orizzontale di dimensioni assai maggiori. Non si sa se questa grande pietra sia mai stata in posizione eretta: le testimonianze più antiche la descrivono come ora si trova.

Un tipo particolare di menhir, noto come statua-menhir, è scolpito in modo da raffigurare una persona, umana o divina. Se ne trovano nella Francia meridionale e nell'Italia settentrionale, e talvolta fino in Spagna e nelle isole della Manica. Queste interessanti pietre erette, le più antiche sculture monumentali a tutto tondo della storia umana, possono essere datate intorno alla fine del III millennio a.C. e all'inizio del II. Non sono, tuttavia, necessariamente da collegare ai menhir privi di decorazione.

La seconda categoria di strutture megalitiche è formata dai gruppi di pietre erette. Le pietre sono poste in allineamenti, come quelli di Carnac, o in quelli che si è soliti definire circoli di pietra. Le accurate ricerche di Alexander Thom, professore di ingegneria a riposo dell'Università di Oxford, hanno rivelato che molti di questi monumenti non sono propriamente circolari, ma presentano la forma di una ellisse o di una ellisse appiattita. Quindi il termine che ora viene preferibilmente adottato è quello di anelli di pietra. Alcuni sorgono isolati; altri circondano un tumulo sepolcrale. Alcuni sono associati ad allineamenti di pietre; altri, come i famosi monumenti di Stonehenge e di Avebury nell'Inghilterra meridionale e di Stennesse Brodgar nelle Orcadi, sono circondati da fossati e argini.

La terza categoria, quella della camera sepolcrale, costituisce la forma più comune di monumento litico europeo. Esistono ancora oggi probabilmente da 40 000 a 50 000 camere di questo genere, e in origine dovevano esservene forse il doppio. Il maggior numero è stato trovato in Spagna, Portogallo, Francia, Inghilterra e Irlanda, Scandinavia e Germania settentrionale. Alcune sono completamente sepolte sotto grandi tumuli di terra e pietrame, altre portano tracce di tumuli che sono andati parzialmente distrutti, e altre sono completamente allo scoperto. Nel XIX secolo si riteneva che tutte le camere allo scoperto fossero ciò che restava di tumuli a camera, ma attualmente tale opinione si è modificata. In qualche caso si tratta effettivamente dei resti di tumuli sepolcrali spogliati della loro copertura, ma altri, come le *portal chambers* (camere con portale) irlandesi e le grandi gallerie in pietra della Francia occidentale, furono probabilmente sempre proprio come le vediamo oggi.

Molte camere sepolcrali presentano una tecnica di copertura più elaborata di quella rappresentata da una semplice pietra appoggiata orizzontalmente; le pietre furono poggiate sugli ortostati in modo tale da sovrapporsi aggettando fino a formare un tetto ricurvo o volta. Tra le più famose volte ad aggetto citiamo Maes Howe nelle Orcadi e Newgrange nella contea di Meath nell'Eire. Maes Howe subì una devastazione da parte dei vichinghi, ma Newgrange, con la sua volta che si eleva da terra fino a circa 6 metri di altezza, si è conservata intatta per circa 4500 anni. È una delle meraviglie dell'Europa preistorica.

Molte delle pietre che compongono queste camere sono di dimensioni assai grandi. Si è calcolato che la pietra di copertura di Mount Browne nella contea di Carlow nell'Eire pesi circa 100 tonnellate. La grande camera megalitica di Bagneux, nei pressi di Saumur nella Francia occidentale, ha una lunghezza di circa 20 metri per una larghezza di 5, e la sua copertura, formata da quattro lastre di pietra, offre tre metri di spazio in verticale a chi vi stia ritto all'interno. (Una volta ospitava addirittura un caffè!). Le pietre di copertura hanno uno spessore di circa 60 centimetri, e si calcola che la più grande pesi 86 tonnellate.

La quarta categoria, quella dei templi megalitici, è la più limitata. La maggior parte degli esemplari si trova nel Mediterraneo, nell'isola di Malta e nella vicina Gozo. L'applicazione del termine «tempio megalitico» a questi grandi monumenti potrebbe far pensare a un edificio simile nella forma alle strutture a pilastri dell'Egitto dinastico e della Grecia classica. Ma i monumenti maltesi sono assai diversi. Le loro solide pareti sono formate da enormi lastre, la pianta evidenzia i profili di absidi, la copertura era probabilmente costituita da travi e paglia. Sono sicuramente i più imponenti monumenti architettonici preistorici dell'intera area mediterranea.

Non sembra possibile dubitare che queste strutture maltesi fossero dei templi. Non contengono sepolture, né tracce di occupazione domestica, mentre vi sono molti oggetti di culto, comprese delle figurine di una divinità femminile, e delle pietre decorate con spirali e altri motivi ornamentali. È anche certa la loro datazione, che si fa risalire intorno al quarto millennio a.C. Si tratta quindi del più antico modello di architettura in pietra esistente nel Mediterraneo e, insieme alle camere megalitiche di Spagna e Francia, nel mondo intero.

Durante il Medioevo sia la gente comune sia gli uomini di cultura ritenevano che i grandi monumenti in pietra dovessero essere stati costruiti da giganti in tempi assai remoti. Effettivamente i primi scavi di cui esista testimonianza, relativi a una camera megalitica in Svezia, furono intrapresi per convalidare tale supposizione. Essa risultò errata: i resti trovati in quella camera non appartenevano a dei giganti. In seguito gli antiquari cercarono di spiegare il fenomeno megalitico come una testimonianza storica scritta. Le strutture furono attribuite con vari criteri ai romani, agli anglosassoni, ai danesi, ai

Questa fotografia, ripresa dal basso verso l'alto, della volta aggettante di Newgrange ne rivela la costruzione del tipo a pietre sovrapposte. Ciascun piano successivo di lastre era collocato in modo tale da restringere sempre più l'apertura, finché era possibile sovrapporvi le pietre di volta. Un tipo analogo di volta aggettante venne costruito a Maes Howe nelle isole Orcadi.

Dolmen, o «tavola di pietra», è il nome di questa struttura megalitica in Bretagna: costituito da un'unica lastra di copertura poggiata su tre o più supporti litici verticali, è la più semplice camera sepolcrale. Il monumento, situato nel Pembrokeshire (Galles) è noto come *cromlech* di Carreg Samson, un termine gallese che equivale a dolmen. Alcune camere sorgono all'aperto, altre sono ricoperte da un tumulo.

Una camera sepolcrale più vasta, nel Pembrokeshire, ora ha soltanto una lastra di copertura. Si trova all'aperto a Pentre-Ifan, e fu citata per la prima volta da George Owen nel suo lavoro di ricerca sulla regione, pubblicato nel 1603. Forse in Europa ne esistono ancora circa 50 000.

goti, agli unni. Per esempio, Herbert, colui che aveva coniato il termine di megalito, sostenne che, siccome Stonehenge e gli altri megaliti dell'Inghilterra non erano menzionati dagli storici romani e dai visitatori della regione, quelle strutture dovevano appartenere a epoca post-romana.

Gli autori e le finalità

Gradualmente si pervenne ad accettare l'idea che i megaliti fossero di epoca pre-romana. Gli abitanti pre-romani dell'Europa nord-occidentale, in particolare gli antichi galli e i bretoni, erano stati descritti da Cesare e da altri scrittori, i quali riferivano che tra quelle genti esisteva una casta di uomini sapienti chiamati druidi. Alcuni inoltre credettero di ravvisare nelle camere all'aperto degli altari druidici.

Negli ultimi due secoli lo sviluppo degli scavi archeologici ha rivelato che le camere megalitiche furono essenzialmente luoghi sepolcrali, usati collettivamente da una comunità o da una famiglia lungo un certo periodo di tempo. Alcune camere, fra le più grandi, hanno fornito i resti di oltre 200 individui. Non tutte le camere megalitiche però furono necessariamente delle tombe. Alcune, come le camere allo scoperto dell'Irlanda e della valle della Loira in Francia, potrebbero essere state dei templi, ma con scarsa probabilità druidici. In realtà, la differenza fra una tomba e un tempio non è chiaramente definibile, come possiamo constatare osservando le chiese e le cattedrali cristiane.

I beni funerari rinvenuti in molte camere megalitiche sono poco appariscenti: vasi privi di decorazioni e asce di pietra levigata. In alcune aree, tuttavia, essi sono ricchi e interessanti. In Scandinavia sono stati rinvenuti vasi fittamente decorati con «oculi», cioè motivi a forma di occhio. In Spagna molte tombe contengono un gran numero di pietre di scisto ricoperte da disegni, alcuni dei quali antropomorfici. In Bretagna le tombe hanno fornito asce levigate di ottima fattura e anelli di giada, granati e altri materiali preziosi. Questi materiali appartengono al III millennio a.C., cioè a un'epoca assai anteriore all'uso della giada da parte dei cinesi e dei maya.

La finalità dei menhir è più difficile da interpretare. Talvolta alla loro base si trovano delle sepolture, ma non possiamo definirli tombe. Inoltre, poiché sepolture di quel tipo non consentono di datare il monumento, tanto meno possono rivelarci il loro scopo originario. Probabilmente queste pietre erette isolate avevano il valore di indicatori territoriali. Un'ipotesi alternativa le considera come monumenti commemorativi per i defunti, una versione preistorica del cenotafio greco, o tomba vuota.

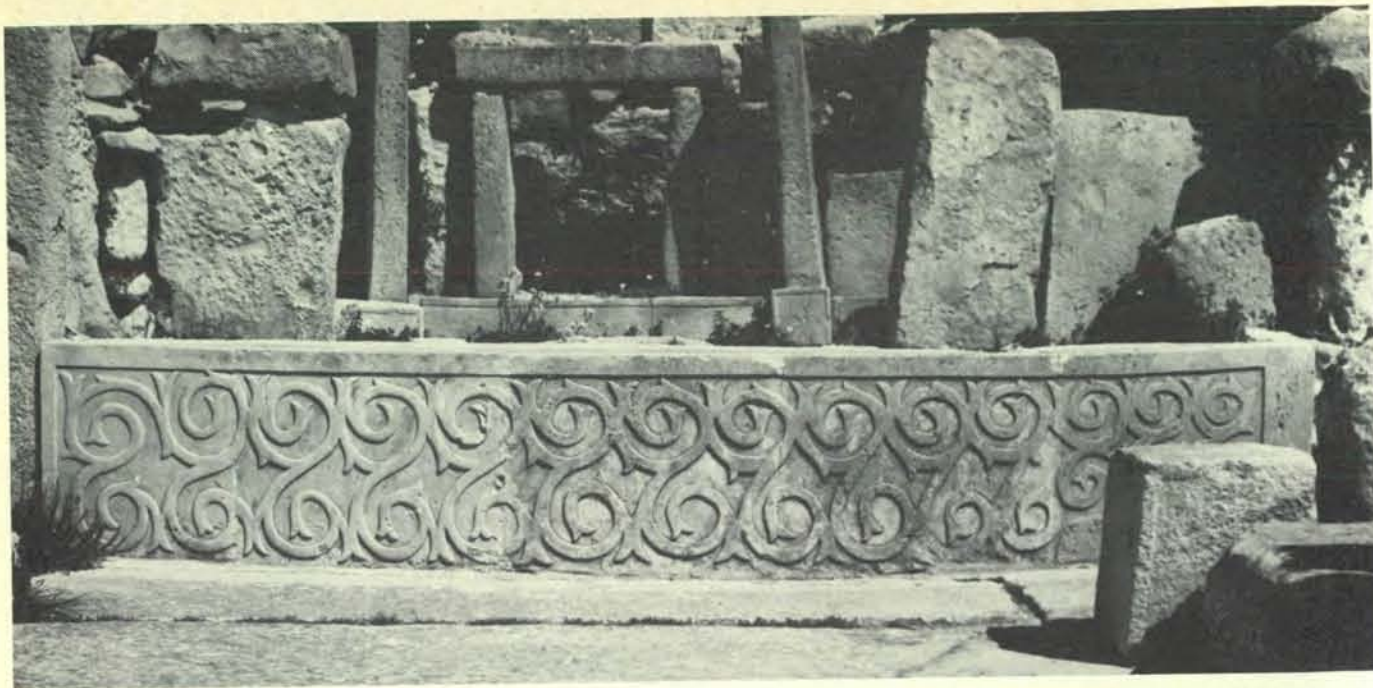
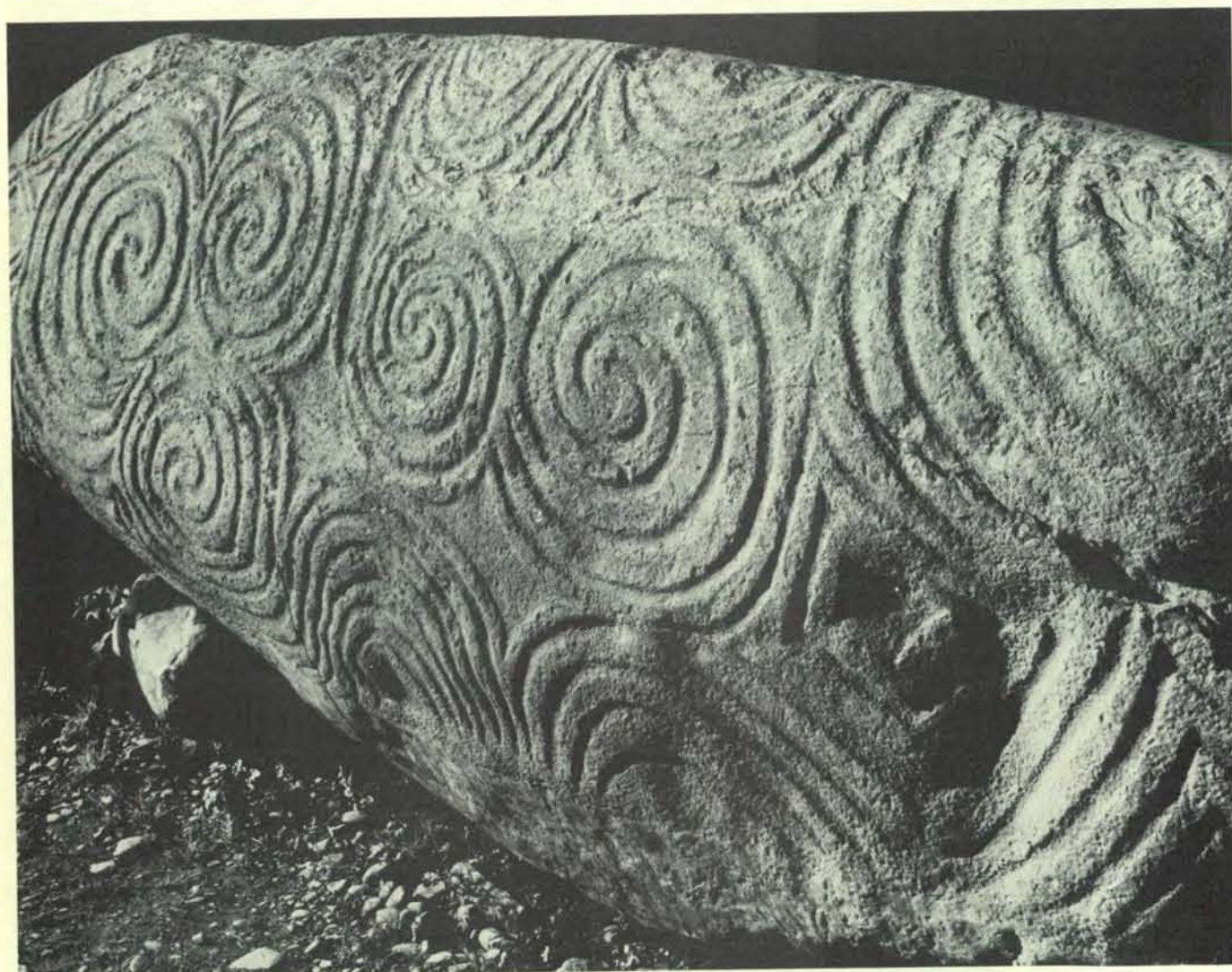
Quanto alla finalità degli allineamenti di menhir e degli anelli di pietra, essa chiaramente non è funeraria né domestica. Probabilmente questi monumenti furono luoghi destinati ad assemblee, nei quali le tribù o le comunità si incontravano di tanto in tanto per motivi di carattere civile o religioso. Io li immagino nella funzione dei *pardons* bretoni dei nostri giorni: sedi appropriate per officiare sacri riti, per benedire raccolti e animali, per incontrare amici e parenti, per partecipare a occasionali eventi di carattere sociale, come mercati e fiere.

Possiamo affermare che una descrizione del genere si adatti a Stonehenge, l'anello di pietra più famoso di tutti? Anzitutto occorre ricordare che Stonehenge è un monumento complesso, passato attraverso molte fasi e con molte caratteristiche architettoniche, tra le quali due circoli, due strutture a ferro di cavallo e dei triliti ben sagomati, ben diversi dai semplici ortostati e architravi. Le pietre stesse sono di materiali differenti; in parte arenarie provenienti dalle Marlborough Downs, e in parte pietre blu «straniere» provenienti dalle colline di Prescelly del Pembrokeshire, nel Galles meridionale. Le pietre blu sono un'ottantina, del peso di quattro tonnellate ciascuna; la distanza da Prescelly a Stonehenge è di circa 217 chilometri in linea d'aria. Anche i massi di arenaria, meglio noti come pietre sarsen, sono di grandi dimensioni e furono accuratamente lavorati. L'elemento più grande dei triliti di sarsen misura circa 9 metri, e l'altezza media delle pietre nel circolo di sarsen è di 4,10 metri. Uno degli accorgimenti più interessanti e architettonicamente sofisticati adottati dai costruttori di Stonehenge è quello noto nell'architettura greca come «entasis». Esso consiste nel dare ai pilastri una forma tale da cancellare l'effetto sfuggente della prospettiva: quando si guardano le pietre da breve distanza esse non appaiono rastremate verso l'alto, ma danno l'illusione ottica di essere perfettamente diritte.

L'erosione o qualche azione vandalica avevano spogliato del tumulo che un tempo la ricopriva questa tomba a camera, quando l'artista ritrasse il monumento intorno al 1840. Solo una lastra orizzontale ha parte della copertura di pietrisco e di terra. L'incisione da cui è stato ricavato lo schizzo di Bryn Celli Ddu ad Anglesey, nel Galles, apparve su *Archaelogia Cambrensis* nel 1847.

La misurazione dell'anno

Non c'è dubbio che Stonehenge fu costruito in modo che il suo asse centrale fosse orientato verso il sorgere del sole di mezza estate. Molti anni fa l'astronomo inglese Sir Norman Lockyer affermò che, calcolando l'esatto orientamento di Stonehenge, egli poteva attribuire una data esatta alla costruzione del monumento. Concluse quindi che tale data era il 1680 a.C. (± 200 anni). In seguito R. J. C. Atkinson dell'University College di Cardiff, che più di ogni altro ha contribuito con i suoi scavi e le sue ricerche alla nostra attuale conoscenza di Stonehenge, ha fatto rilevare che, a causa di errori contenuti nelle originali argomentazioni di Lockyer, quella data doveva essere leggermente modificata. La nuova lettura è 1840 a.C. (± 200 anni).

La finalità astronomica di Stonehenge e degli altri anelli di pietra non fu considerata a fondo che dopo la seconda guerra mondiale, quando Gerald S. Hawkins della Boston University nel suo libro *Stonehenge Decodec (Decodificazione di Stonehenge,* 1966) avanzò l'ipotesi che il monumento fosse un gigantesco calcolatore per la predizione delle eclissi lunari e solari. Cinque anni più tardi Thom, nel suo libro *Megalithic Lunar Observatories (Osservatori lunari megalitici),* suppose che molti monumenti megalitici servissero all'osservazione dei moti e delle fasi lunari. Gli studi di Thom lo avevano poi indotto a sostenere l'esistenza di una «yard» megalitica equivalente a circa 0,8 metri e a ipotizzare che i costruttori degli anelli di pietra avessero conoscenza della geometria pitagorica 2000 anni o più prima dei greci. Sono affermazioni stravaganti e poco convincenti: quello che

Motivi a doppia spirale che decorano una bassa parete nel tempio maltese di Hal Tarxien, nei pressi di La Valletta. I numerosi templi megalitici di Malta e della vicina Gozo, un tempo forse ispirati da strutture minoiche, sono ora riconosciuti come forme di origine indigena.

La decorazione a spirale è il motivo predominante fra quelli eseguiti in bassorilievo su questa colonna megalitica all'ingresso di Newgrange. La decorazione se risale alla stessa epoca della camera sepolcrale è almeno di 1500 anni più tarda dei templi megalitici più antichi di Malta.

possedevano i costruttori di megaliti era la capacità pratica di tracciare triangoli rettangoli.

Molta gente, senza dubbio tediata dalla prosaica descrizione dei megaliti risultante dalla ricerca archeologica, accolse con interesse le tesi affascinanti di Hawkins e di Thom, accettando l'idea che i costruttori di megaliti fossero non soltanto esperti di geometria pitagorica e possessori di precise unità di misura, ma anche astronomi provetti che si dedicavano allo studio delle eclissi, dei moti lunari e della posizione delle stelle. Il che equivale, secondo me, a una specie di elegante versione accademica dell'archeologia astronautica. Gli appassionati di archeoastronomia, sebbene molto opportunamente respingano l'idea di uomini superiori provenienti dallo spazio esterno, altrettanto inopportunamente insistono sulla presenza nell'Europa antica di uomini colti con una passione per l'astronomia di ispirazione evidentemente religiosa. Secondo la mia opinione, l'interpretazione dei monumenti megalitici come osservatori astronomici non è mai stata suffragata da prove. Tale interpretazione appare soggettiva e imposta dall'osservatore. Nuovi studi stanno già dimostrando l'imprecisione di alcune delle precedenti osservazioni e vanno indebolendo le speranze di coloro i quali ritengono che i costruttori di megaliti fossero legati a un culto astronomico.

L'intero studio dei megaliti in Europa è stato rivoluzionato non dai ricercatori che volgevano lo sguardo alla Luna e alle stelle, ma dai progressi compiuti nel campo della datazione preistorica. La prima di queste nuove tecniche, quella di Willard F. Libby e dei suoi colleghi, è la datazione con il carbonio-14. Le prime date ottenute con tale metodo furono pubblicate 30 anni fa, e da allora ne sono state determinate a migliaia nei laboratori di tutto il mondo. Una tecnica geocronologica ancora più recente, la datazione basata sulla termoluminescenza, ha confermato molti dei dati ottenuti con il carbonio-14.

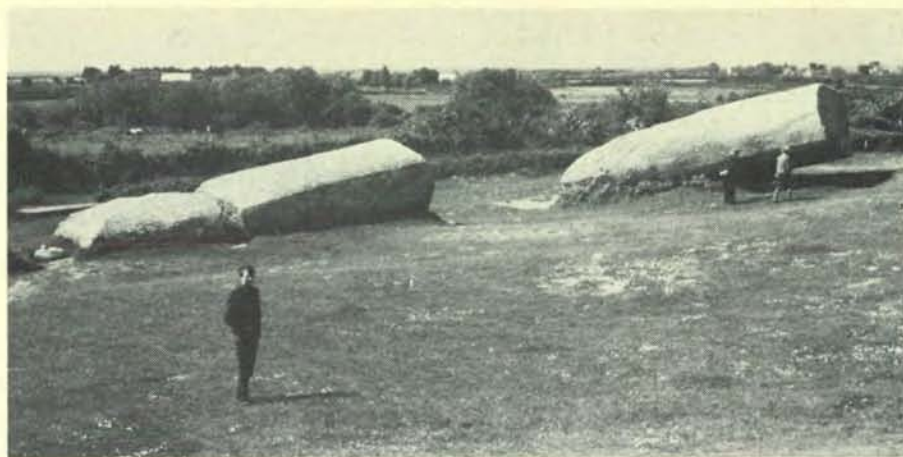
Il problema della datazione

Prima della rivoluzione provocata dal carbonio-14 - non si può usare un termine meno forte se si considerano gli effetti prodotti nel campo dell'archeologia preistorica - le età delle diverse antiche attività dell'uomo erano state ipotizzate o calcolate in modo deplorevolmente incerto, in base a correlazioni tra le civiltà datate dell'Egitto e della Mesopotamia e l'Europa ancora barbarica e priva di datazione. Nel periodo, quindi, compreso tra le due guerre mondiali, era consuetudine attribuire i più antichi megaliti europei al Neolitico, cioè più o meno a un arco di tempo compreso fra il 2500 e il 2000 a.C., e i grandi monumenti, come Stonehenge e Newgrange, all'età del bronzo, tra il 1800 e il 1500 a.C.

Ora, grazie al metodo rivoluzionario del carbonio-14, possiamo con fiducia stabilire l'effettiva età dei megaliti europei. I templi maltesi sono da attribuire a un periodo che si estende dal 4000 al 2000 a.C. Le tombe megalitiche di Spagna e Portogallo risalgono a un periodo compreso fra il 3800 e il 2000 a.C. Le tombe britanniche e irlandesi si datano da poco dopo il 3800 fino al 2000 a.C., e quelle scandinave da poco prima del 3000 al 1800 a.C. circa. Nella sequenza irlandese la datazione ottenuta per Newgrange ha consentito di dimostrare che la sua costruzione risale al 2500 a.C. circa. Nella sequenza britannica Atkinson ha stabilito i particolari cronologici di Stonohenge: una prima fase, dal 2800 al 2200 a.C.; una seconda, che comprende l'arrivo dal Galles delle pietre blu, dal 2100 al 2000 a.C.; una terza, che comprende tre fasi in subordine, dal 2000 al 1100 a.C, epoca in cui ebbe inizio la fase finale.

Le date da me citate per Stonehenge e per i monumenti megalitici in generale sono quelle definite come date calibrate del carbonio-14. Le date, cioè, sono state modificate secondo le correzioni conseguenti allo studio degli anelli del *Pinus aristata,* quali appaiono nelle curve di variazione dei diagrammi tracciati da Hans E. Suess dell'Università della California di San Diego, da R. Malcolm Clark della Monash University in Australia e da altri.

Ora che sono stati datati con precisione, come spiegare le origini di questi monumenti megalitici? Per quanto riguarda i menhir della Bretagna e di altre

Un motivo semicircolare predomina nelle sculture in bassorilievo di queste pietre verticali della camera sepolcrale megalitica di Gavrinis in Francia, contemporanea all'incirca di Newgrange. Non sappiamo se i disegni erano decorativi o se avevano un qualche significato religioso.

Il più grande menhir dell'Europa occidentale è questo esemplare spezzato di Locmariaquer in Bretagna. In questa fotografia sono visibili tre dei suoi quattro frammenti; riuniti insieme misurerebbero 22 metri di lunghezza. Non siamo in grado di dire se il menhir sia mai stato in posizione verticale; le testimonianze più antiche lo indicano sempre nella sua attuale posizione.

Un menhir cristianizzato a Rudston, nello Yorkshire. Con la sua lunghezza di 7,5 metri, è il più alto in Inghilterra. Il monumento megalitico non fu abbattuto; il cimitero gli si sviluppò attorno.

zone dell'Europa nord-occidentale esiste una sola spiegazione valida: si tratta di una invenzione locale. Essi rappresentano probabilmente il trasferimento nella pietra di una pratica più antica, consistente nell'erigere pilastri di legno come cenotafi o come indicatori territoriali, oppure come totem.

Anche per i templi maltesi attualmente abbiamo una sola plausibile spiegazione. Si era soliti affermare che si trattava di strutture derivate dalla diffusione della civiltà minoica di Creta o di Micene, o di luoghi ancora più lontani. La cronologia con il carbonio-14 rivela ora che essi sono più antichi di qualsiasi contesto minoico, miceneo, egiziano o sumerico. I templi maltesi costituirono un fenomeno indigeno. Probabilmente sono versioni in superficie delle tombe-templi sotterranee scavate nella roccia. In ogni caso essi non sembrano avere precedenti in altre regioni e non sono mai state ritrovate, al di fuori di Malta e di Gozo, strutture che si possano attendibilmente far risalire ai templi maltesi.

Veniamo ora alle origini delle camere megalitiche e degli anelli di pietra. Essi non sono da considerare fenomeni isolati. Dobbiamo invece cercare di spiegarli nel contesto delle società del Neolitico che li crearono. L'ipotesi di Elliot Smith che i megaliti europei fossero una derivazione della «mastaba», o tomba di pietra dell'Egitto, non ha fondamento nei fatti né indizi di probabilità. Sappiamo ora infatti che i megaliti europei sono più antichi delle mastaba e delle piramidi egizie.

Un'altra ipotesi ancora, espressa da V. Gordon Childe e da altri, e tenacemente sostenuta dalla maggior parte degli archeologi durante gli anni dal 1925 al 1950, asseriva che i megaliti europei erano stati eretti da popoli originari del Mediterraneo orientale, in particolare di Creta e delle isole dell'Egeo. Anche prima dell'introduzione della datazione col carbonio-14, si andava chiarendo comunque l'idea che, per esempio, era cronologicamente impossibile far derivare le grandi tombe a camera come quelle di Newgrange e di Maes Howe dalle tombe a cupola sotterranee di Micene. Una seconda linea di difesa di tale ipotesi consistette nel retrodatarla, cercando una connessione con le tombe a volta di Messara a Creta e le tombe scavate nella roccia delle Cicladi. Ora, sia le determinazioni col carbonio-14 che quelle ottenute con la termoluminescenza dimostrano che le camere megalitiche europee sono di età anteriore a tutte le tombe collettive del Mediterraneo orientale.

La nuova conclusione

Siamo perciò obbligati a concludere che le tombe a camera ebbero un'origine indipendente in almeno sette diverse zone d'Europa: Spagna meridionale, Portogallo, Bretagna, Francia settentrionale, Germania settentrionale, Scandinavia, Inghilterra meridionale e Scozia. Questa

asserzione costituisce dunque una vittoria della scuola di teoria archeologica che sostiene l'«evoluzione indipendente» sulla scuola «diffusionista» che ha tenuto il campo tanto a lungo?

La risposta non è così semplice come può apparire. Le società del Neolitico che incominciarono a erigere megaliti in alcune zone europee intorno al 4000 a.C. erano quelle che già costruivano strutture analoghe non litiche o non megalitiche, cioè abitazioni e tumuli. Molti di noi ritengono attualmente che una spiegazione alla comparsa delle tombe a camera megalitiche in molte e diverse parti d'Europa deve prendere in considerazione tre fasi successive. La prima fase postula l'esistenza di una antica tradizione neolitica europea per la costruzione di abitazioni in legno e pietra. La seconda comporta la trasformazione di queste strutture domestiche in tombe, ancora costruite in legno, torba o pietre non megalitiche. La terza fase vede il trasferimento di queste tradizioni neolitiche di ampia diffusione in una architettura di tipo megalitico, in aree diverse dell'Europa che senza dubbio erano collegate fra loro.

L'origine degli anelli di pietra può essere spiegata per gran parte allo stesso modo, sebbene in questo caso il fenomeno non riguardi sette aree diverse ma interconnesse. Gli anelli infatti si trovano soltanto in Inghilterra e Irlanda. Come la tomba a camera presuppone l'esistenza di un più semplice precursore, così gli anelli di pietra non possono essere dissociati dai monumenti di legno ad anello, comunemente indicati col nome di *woodhenge*.

Secondo la mia opinione gli anelli di pietra ebbero origine nel modo seguente: da principio esistevano delle radure circolari nelle foreste che ricoprivano l'Europa neolitica nel V e IV millennio a.C. È lecito supporre che in queste radure avessero luogo delle adunanze, sia di tipo religioso che laico. In un secondo tempo, per effetto della presenza degli animali domestici e dell'uomo stesso, la foresta scomparve, dopodiché le radure furono create artificialmente piantando dei pali in circolo per formare una specie di piattaforma ove tenere simili adunanze. La terza fase consistette nel passaggio dagli anelli di legno a quelli di pietra. Infine poi, come suprema fatica di una serie di uomini che possiamo soltanto definire architetti di cattedrali, verso la metà del III millennio a.C. fu costruito Stonehenge, che fiorì come tempio - luogo di adunanze - stadio per oltre mille anni. Questo ci riporta al precedente interrogativo: si adatta Stonehenge alla descrizione degli anelli megalitici in generale come luoghi di incontri sacri e profani? La risposta sembra essere nettamente affermativa.

I megaliti e la religione

È facile intuire che dovette essere la forza di una fede a ispirare le popolazioni del Neolitico che profusero le loro fatiche nella costruzione dei megaliti europei. Tombe, templi, cenotafi, luoghi di riunione - qualunque cosa fossero i monumenti - ci appaiono come le manifestazioni di qualche potente credenza religiosa, che comprendeva anche la fede in una vita d'oltretomba. Era tale religione collegata all'annuale percorso del Sole, come lascia supporre l'orientamento di Stonehenge e degli altri anelli di pietra? Era collegata a una dea madre, come sembrano suggerire le figurazioni dei templi maltesi e quelle sul vasellame e sulle lastre di scisto delle tombe di Spagna e Portogallo? Non lo sapremo probabilmente mai. Senza fonti scritte è praticamente impossibile ricostruire le caratteristiche, religiose e sociali, delle società preistoriche. Non conosciamo il motivo che spinse i cacciatori del Paleolitico superiore a creare pitture, incisioni e sculture nella Francia meridionale e nella Spagna settentrionale, e dobbiamo parimenti ammettere che non conosciamo l'etica sociale e religiosa dei costruttori di megaliti.

Il recinto di Avebury, riconosciuto per la prima volta come circolo megalitico da John Aubrey nel XVII secolo, indusse costui a far rilevare che «soverchia in grandezza il famoso Stonehenge quanto una cattedrale soverchia una chiesa parrocchiale». In questa fotografia aerea sono visibili alcuni dei 190 menhir disuguali che formano il circolo e i suoi viali, ma l'argine e il fossato risultanti dallo scavo di 200 000 tonnellate di materiale calcareo racchiudono più di 11 ettari. Il tumulo conico in alto a destra, chiamato Silbury Hill, è pure opera dell'uomo.

Alcune tra le migliaia di tombe megalitiche europee sono decorate con motivi incisi in bassorilievo o scolpiti sulla superficie della pietra. Nella Germania centrale alcuni sembrano riproduzioni di arazzi, probabilmente tappezzerie che rivestivano le pareti delle case. Altri, specialmente nel bacino di Parigi, sono chiare rappresentazioni di una figura femminile vista di fronte: sono in evidenza gli occhi, il naso e i seni della dea. La maggioranza dei disegni megalitici nelle tombe bretoni e irlandesi, tuttavia, è costituita da schemi geometrici: spirali, zig-zag, losanghe, cerchi concentrici. Ritroviamo forse i più belli a Gavrinis in Bretagna e a Newgrange. Alcuni, come le spirali che vediamo sulla grande pietra all'ingresso di Newgrange, sembrano semplici motivi decorativi; altri si presentano ai nostri occhi in una sconcertante confusione di probabili segni o simboli. È importante ricordare che alcune di queste decorazioni geometriche sono sepolte nei tumuli che racchiudono le tombe a camera e non furono mai destinate a essere viste: esse furono deliberatamente celate all'epoca della costruzione. Per quale motivo? Sono messaggi o simboli sacri? Qualunque possa essere il loro significato esso è stato posto in un carattere che non sapremo mai decifrare.

A questo proposito, soltanto 15 anni fa si è scoperto che su uno dei pilastri di Stonehenge era stato inciso il disegno di un pugnale, oltre a molte raffigurazioni di asce, nello stile dell'antica età del bronzo. Alcuni supposero che si trattasse di un pugnale miceneo. Questo genere di arte è assai diverso da quello delle tombe a camera, e naturalmente queste incisioni possono essere state eseguite in epoca tarda nel corso della storia di Stonehenge. Dire che si tratta di graffiti dell'età del bronzo, tuttavia, non significa che in quell'età il monumento fosse considerato alla stregua di una gigantesca ritirata.

La consuetudine di costruire grandi monumenti in pietra ebbe termine in Europa intorno al 1000 a.C., ma non se ne trae necessariamente la conclusione che le popolazioni dimenticassero simultaneamente la natura e il significato di quelle strutture. Non è affatto improbabile che tutte le popolazioni del I millennio a.C. continuassero a frequentare gli anelli di pietra come luoghi di raduno e di culto, ed è anche più che possibile che la classe sacerdotale druidica dei celti pre-romani della Gallia e della Britannia li usasse come templi. Non abbiamo, tuttavia, alcuna testimonianza archeologica in proposito.

Da principio la cristianità disapprovò fortemente le genti che adoravano le pietre, ma gradatamente si pervenne a un nuovo spirito di tolleranza, che risultò tanto generoso da portare alla cristianizzazione di alcuni menhir. In Spagna e in Bretagna, infatti, alcuni monumenti megalitici risultano associati a moderne chiese cristiane in attività. Io considero questo fatto come un segno che l'antica fede dei costruttori sopravvisse in qualche modo o forma almeno fino al Medioevo dell'Europa occidentale.

La perizia dei costruttori

Ci si è spesso domandati chi fossero i costruttori dei megaliti e dove fossero le loro abitazioni e i loro insediamenti. Noi riteniamo però che questa ricerca debba procedere nella direzione opposta; sono gli abitanti dei villaggi neolitici dell'Europa del IV e III millenio a.C. che in certe aree costruirono le loro tombe e i loro templi in solida pietra. Molto è stato detto e scritto in tempi recenti sulla rivoluzione verificatasi nella nostra visione del passato preistorico. Agli archeologi si rimprovera di aver considerato le popolazioni neolitiche come selvagge, mentre il nuovo modo di vedere il passato, che di quelle rivela la grande perizia e il possesso di conoscenze matematiche, geometriche e astronomiche, muta profondamente l'immagine che ci eravamo costruita dell'uomo di quei tempi.

Questa tesi può trovare impreparati soltanto coloro che non hanno mai compreso il valore della testimonianza archeologica e vorrebbero rendere sensazionale la preistoria. Per quanto mi riguarda, non ho mai dubitato che le genti del Neolitico europeo avessero buone capacità tecniche e ingegneristiche. Esse seppero scavare pietre di grandi dimensioni, le trasportarono a distanze considerevoli e le eressero con arte e abilità consumata. Possiamo avere un'idea circa le probabili tecniche da esse impiegate studiando i metodi attualmente in uso in certe regioni, come l'Assam, ove ancora si costruiscono monumenti megalitici. Possiamo avere una misura delle loro imprese rilevando le difficoltà che attualmente incontrano gli agricoltori e gli impresari edili quando cercano di fare a pezzi le strutture preistoriche. Ad esempio, verso la fine del secolo scorso la pietra di volta di una camera megalitica nei pressi di Saumur fu rimossa per impiegarla come ponte attraverso un corso d'acqua locale. Gli addetti all'opera di rimozione prepararono un certo numero di enormi rulli di oltre un metro di circonferenza: ogni rullo era stato ottenuto legando insieme i tronchi di quattro querce. Anche con l'impiego dei rulli fu necessario far ricorso a 18 paia di buoi per muovere quel carico.

Non dovremmo quindi mai disconoscere la grandezza delle imprese affrontate dai costruttori megalitici e neppure negare che, di quando in quando, durante le pause dalle loro fatiche nei campi o durante la costruzione di un monumento, essi abbiano rivolto come tutti noi lo sguardo verso il Sole, la Luna, le stelle. Sono trascorsi 6000 anni da quando a Malta o in Bretagna fu eretto il primo megalito; 4000 anni prima dell'era cristiana e 1000 anni prima delle civiltà letterate di Egitto e Mesopotamia. È una modesta riflessione che possiamo fare quando, con orgoglio e piacere, contempliamo il nostro patrimonio megalitico.

Il lupus eritematoso sistemico

Questa malattia che colpisce vari organi è legata a una disfunzione del sistema immunitario. Il paziente produce autoanticorpi che si legano agli antigeni formando immunocomplessi dannosi verso i tessuti

di David Koffler

Una delle prime manifestazioni del lupus eritematoso sistemico (LES) è un'infiammazione erosiva della pelle che compare tipicamente sotto forma di chiazze convergenti sul naso e sulle guance e che ha dato a questa malattia il suo nome curioso: *lupus eritematosus*, ossia «lupo rosso». Questo nome ovviamente non riflette la vera portata e la natura multisistemica di questa malattia reumatoide che colpisce negli Stati Uniti una donna su 500 (e gli uomini con un frequenza dieci volte minore). Si tratta di una malattia sistemica, che colpisce non solo la pelle ma anche, in pazienti e in tempi differenti, le articolazioni, i vasi sanguigni, il cuore, i polmoni, il cervello e, in modo molto grave, i reni.

Anche se i diversi pazienti presentano sintomi differenti, questa malattia è essenzialmente una disfunzione del sistema immunitario. Per un certo periodo, il LES è stato considerato una «malattia autoimmune», cioè una malattia in cui il paziente produce anticorpi che interagiscono con gli antigeni dei propri tessuti. Sebbene ciò possa verificarsi in alcune manifestazioni della malattia, si ritiene oggi che il LES sia principalmente una malattia dovuta ai complessi antigene-anticorpo: gli anticorpi del paziente si combinano con i loro antigeni (non è ancora noto se si tratti di antigeni del paziente stesso o estranei) e formano complessi che sono le vere cause del danno ai tessuti.

La funzione principale del sistema immunitario è di distinguere il «proprio» *(self)* dall'«estraneo» *(non-self)* ossia ciò che fa parte dell'organismo da ciò che proviene dall'esterno. In questo modo l'ambiente interno dell'organismo viene protetto dalle infezioni e dagli altri pericoli provenienti dal mondo esterno. Nell'uomo il sistema immunitario consiste di forse mille miliardi (10^{12}) di cellule chiamate linfociti e di circa 10^{20} molecole di immunoglobuline. Esistono due tipi di linfociti: le cellule *B* e le cellule *T*. Ogni clone o linea cellulare di cellule *B* è programmato geneticamente per produrre un anticorpo particolare con la capacità di riconoscere e combinarsi con un particolare epitopo (la regione di una molecola di antigene con cui l'anticorpo interagisce).

Nel suo stato di riposo un linfocita *B* porta una parte degli anticorpi che produce sulla propria membrana cellulare sotto forma di recettori. Il contatto di un recettore con il suo antigene specifico stimola la proliferazione e il differenziamento della cellula *B*, che dà origine a un clone di plasmacellule che producono milioni di molecole identiche di anticorpo. Ci sono cinque classi principali di anticorpi, due delle quali svolgono un ruolo di primaria importanza nel LES e in altre malattie reumatiche: le immunoglobuline (o gammaglobuline) IgG e IgM. Gli anticorpi si combinano nel sangue o nei tessuti con l'antigene e formano in questo modo gli immunocomplessi (complessi antigene-anticorpo), oltre a innescare vari processi che possono distruggere le cellule portatrici dell'antigene o che possono danneggiare gli spettatori innocenti: i tessuti dell'organismo stesso.

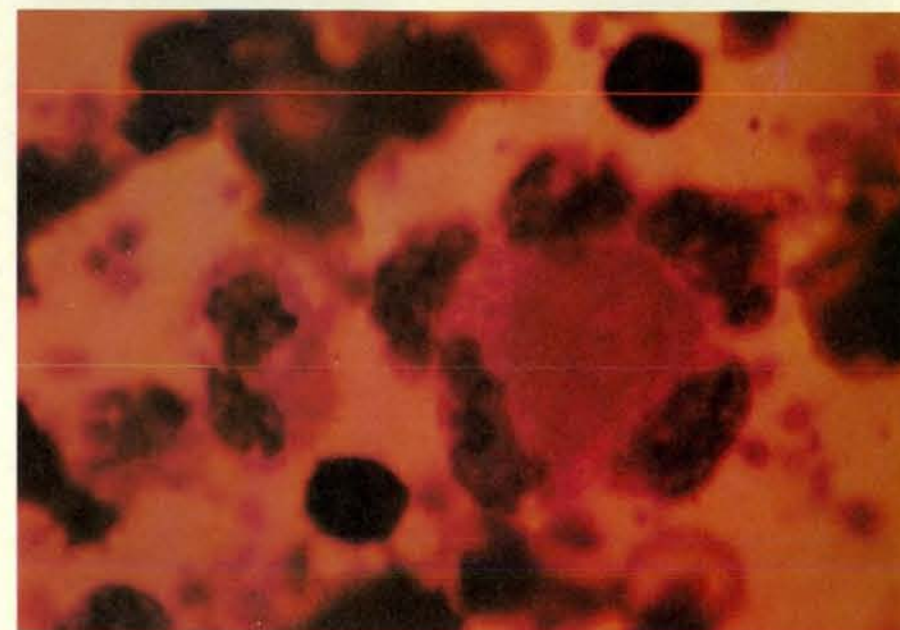
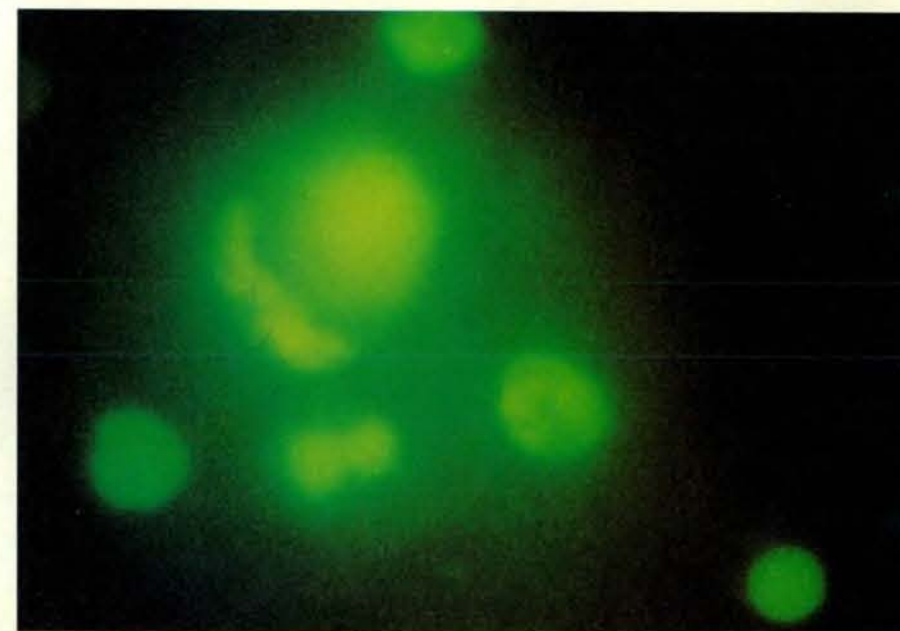
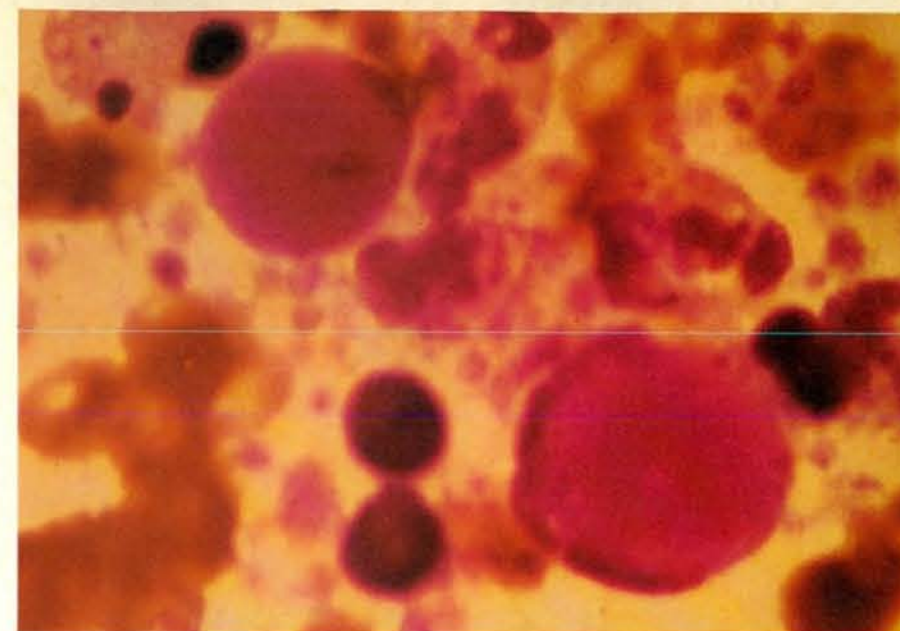
I più importanti immunocomplessi sembrano essere quelli che fissano il complemento, un gruppo di nove proteine del siero. Una porzione della prima componente del complemento, designata C1q, interagisce con la parte anticorpale del complesso e dà inizio a una serie di reazioni che producono una risposta infiammatoria. Alcuni fattori formati dal complemento possono legarsi ai neutrofili (i globuli bianchi più comuni), stimolandone la fagocitosi (ingestione di particelle) e il rilascio di enzimi che demoliscono le proteine. Piccole molecole derivate dal complemento attraggono altre cellule capaci di fagocitosi; altri fattori aumentano la permeabilità dei vasi sanguigni inducendo il rilascio di istamina da parte di alcune cellule. L'afflusso di cellule infiammatorie viene aumentato, così come la deposizione di immunocomplessi.

Mentre i linfociti *B* sono i mediatori della risposta immunitaria umorale che riguarda gli anticorpi solubili, i linfociti *T* sono i mediatori dell'immunità cellulare. Anch'essi possiedono sulla loro superficie recettori specifici che riconoscono gli epitopi sulle molecole di antigene. Il riconoscimento di un antigene induce il rilascio da parte delle cellule *T* delle linfochine, sostanze chimiche che hanno molti effetti: danneggiano le cellule, inibiscono la replicazione dei virus e attraggono le cellule infiammatorie. Altri linfociti *T* influenzano la produzione di anticorpi da parte delle cellule *B*. Un tipo, le cellule dette *helper* (facilitatrici), stimolano la proliferazione delle cellule *B*, mentre un altro tipo, le cellule dette soppressive, la ritardano. Sembra che l'importanza delle cellule *T* per il LES sia dovuta a una alterazione della loro funzione regolatrice piuttosto che a un danno diretto ai tessuti.

Benché il sistema immunitario generalmente sia in grado di distinguere il «proprio» dall'«estraneo» questa distinzione non è affatto assoluta. Già nel 1902 Paul Ehrlich richiamò l'attenzione sulla possibilità che il sistema immunitario producesse danni all'organismo stesso. Ehrlich postulò l'esistenza di un *horror autotoxicus* che escluda questa possibilità: «l'organismo possiede dei meccanismi che impediscono [al sistema immunitario] di reagire contro gli elementi stessi dell'organismo». Tuttavia questo non è vero: ci sono molte situazioni in cui questa risposta autoimmunitaria si verifica. Forse la prima malattia autoimmunitaria o autoimmune provocata involontariamente dai medici è stata un'encefalomielite (infiammazione del midollo spinale e del cervello) indotta nei pazienti che erano stati morsicati da un cane e a cui erano state somministrate parecchie iniezioni del vaccino di Pasteur per prevenire la rabbia. Il vaccino, preparato coltivando il virus della rabbia nel midollo spinale di coniglio, conteneva antigeni che davano una reazione crociata con gli antigeni del sistema nervoso centrale dell'uomo. Ciò significa che gli antigeni del coniglio e dell'uomo hanno almeno alcuni epitopi in comune e che gli anticorpi la cui produzione era stimolata dal vaccino riconoscevano e interagivano con gli epitopi umani.

Nel 1933 Thomas M. Rivers del Rockefeller Institute for Medical Research dimostrò che iniezioni ripetute di un omogenato di tessuto nervoso in alcune scimmie potevano indurre sperimentalmente una encefalomielite simile. Tuttavia fu solo nel 1956 che Ernest Witebsky e Noel R. Rose della University of Buffalo School of Medicine definirono chiaramente le malattie autoimmuni e pubblicarono un metodo riproducibile per abbassare la naturale tolleranza del sistema immunitario verso i propri antigeni. Essi immunizzarono i conigli con estratti di tiroide di coniglio mescolata a un adiutore (una sostanza che incrementa la risposta immunitaria) e i conigli furono colpiti da tiroidite, un'infiammazione della tiroide. Witebsky e Rose definirono i criteri per definire una malattia autoimmune: l'immunizzazione con un antigene come una particolare proteina della tiroide dà inizio a una risposta immunitaria contro l'antigene seguita da alterazioni patologiche limitate all'organo che contiene l'antigene. Più tardi William O. Weigle della Scripps Clinic and Research Foundation di La Jolla, in California, trovò che una tiroidite simile poteva essere indotta immunizzando i conigli con proteine della tiroide di specie diverse in grado di dare una reazione crociata con la proteina della tiroide del coniglio; queste proteine venivano riconosciute come proteine estranee, ma possedevano epitopi identici agli epitopi della tiroide del coniglio. Il meccanismo della malattia autoimmune è quindi l'attacco diretto da parte del sistema immunitario del corpo contro i tessuti stessi dell'organismo e coinvolge probabilmente sia gli anticorpi solubili, sia le cellule *T* sensibilizzate.

Esiste un altro meccanismo «autolesionistico» da parte del sistema immunitario. Clemens von Pirquet della Johns Hopkins University School of Medicine lo riconobbe nei primi anni del novecento studiando le cause della «malattia da siero» riscontrata nei pazienti trattati con un antisiero di cavallo, per esempio per neutralizzare gli effetti della tossina del tetano. Von Pirquet suggerì che la malattia da siero fosse provocata dall'interazione

Il primo test diagnostico per il lupus eritematoso sistemico (LES) era basato sugli anticorpi che inducono la formazione di cellule caratteristiche (cellule LE). I globuli bianchi vengono agitati in una provetta per danneggiarli e incubati con il siero di un paziente. Se il siero contiene gli anticorpi antinucleo caratteristici del LES, questi penetrano nelle cellule danneggiate e provocano il rigonfiamento, la degenerazione e l'espulsione dei loro nuclei sotto forma di corpi nucleari liberi *(sopra, a sinistra nella microfotografia in alto)*. Il corpo nucleare può venire inghiottito da un neutrofilo, con la formazione di una cellula LE *(sotto, a destra nella microfotografia in alto)*; il nucleo del neutrofilo viene compresso verso il bordo della cellula, come dimostra più chiaramente una colorazione per il DNA *(microfotografia al centro)*. Talvolta un corpo nucleare è circondato da una «rosetta» di neutrofili *(in basso)*.

degli anticorpi del paziente con gli antigeni estranei (le proteine del siero di cavallo) con formazione di una sostanza tossica. Aveva ragione: la sostanza tossica è un immunocomplesso responsabile dell'eruzione cutanea, dei dolori articolari e degli altri sintomi della malattia da siero. Il meccanismo del danno ai tessuti in una malattia di questo tipo consiste nella formazione di immunocomplessi tossici in seguito alla risposta immunitaria provocata dall'antigene.

Sia gli autoanticorpi che gli immunocomplessi sono ubiquitari nel siero dei mammiferi. Quasi certamente la loro presenza non è casuale: al contrario sembra che siano componenti essenziali del sistema immunitario e svolgano un ruolo vitale nella conservazione dell'omeostasi, cioè della stabilità dell'ambiente interno dell'organismo. Secondo la teoria della rete funzionale proposta da Niels Kaj Jerne dell'Istituto di immunologia di Basilea, per esempio, una particolare classe di anticorpi, gli anticorpi anti-idiotipici, agiscono nell'autoregolazione del sistema immunitario (si veda l'articolo *Il sistema immunitario*, di Niels Kaj Jerne, in «Le Scienze», n. 63, novembre 1973). Questi anticorpi riconoscono gli epitopi (chiamati idiotipi) che si trovano in corrispondenza o vicino al sito di combinazione delle molecole IgG e sono caratteristici per gli anticorpi prodotti da un singolo clone di plasmacellule. Questi singolari epitopi furono descritti per la prima volta nell'uomo da Henry G. Kunkel della Rockefeller University e negli animali da Jacques Oudin dell'Istituto Pasteur. Gli anticorpi anti-idiotipici sono stati identificati negli animali di laboratorio ed è stato dimostrato che l'induzione di anticorpi anti-idiotipici stimola oppure sopprime la risposta immunitaria dell'animale.

La risposta immunitaria è prodotta dagli anticorpi in circolo e dalle cellule sensibilizzate. Le cellule indifferenziate nel midollo osseo *(1)* si differenziano formando cellule linfocitarie immature che migrano nel tessuto linfoide associato all'intestino *(2a)* e nel timo *(2b)*, dove maturano e formano rispettivamente i linfociti *B* e *T* *(3)*. Una cellula *B* è programmata geneticamente per produrre poche copie di un anticorpo specifico che essa utilizza come recettore sulla propria superficie. Diversi tipi di cellule *T*, con recettori abbastanza simili, hanno ruoli regolativi o vengono sensibilizzate dall'antigene per funzionare nell'immunità cellulare. Esistono cinque classi di anticorpi, di cui IgG e IgM sono le più importanti nel LES. Qui una cellula produttrice di IgM diviene in grado di produrre IgG *(4)*. Un epitopo (una piccola regione su una molecola di antigene) interagisce con il sito di combinazione di un recettore IgG, provocando il differenziamento della cellula *B* in modo che produca un clone di plasmacellule *(5)* che sintetizzano grandi quantità di un particolare anticorpo, che può venire identificato per mezzo del suo idiotipo unico, un epitopo localizzato in corrispondenza del sito di combinazione dell'anticorpo. Lo stesso idiotipo può essere presente sui recettori superficiali delle cellule *T* che regolano la risposta immunitaria. L'anticorpo può reagire con un epitopo sullo stesso antigene che ha originariamente stimolato la proliferazione della cellula *B*, formando in questo modo un immunocomplesso *(6)*.

Nell'uomo uno dei gruppi più comuni di autoanticorpi è costituito dai «fattori reumatoidi», che sono anticorpi diretti contro le gammaglobuline. I fattori reumatoidi sono presenti a bassa concentrazione negli individui normali e in concentrazione molto più alta nei pazienti affetti da varie malattie: sono particolarmente abbondanti nell'artrite reumatoide e nel LES, dove possono svolgere un ruolo importante nel meccanismo della malattia. Altri anticorpi sono diretti contro i materiali del nucleo cellulare; sono presenti in una piccola percentuale di individui normali, con una frequenza leggermente maggiore nelle persone anziane e molto maggiore nei pazienti affetti dal LES, come si vedrà più avanti. (È importante sottolineare il fatto che la distinzione tra un livello anticorpale «normale» e uno «anormale» dipende dalla sensibilità del metodo di dosaggio e diventa sempre più confusa mano a mano che si rendono disponibili metodi diagnostici più precisi.)

Qual è la funzione biologica degli autoanticorpi e degli immunocomplessi normalmente presenti anche in assenza di antigeni estranei? Le indicazioni sinora disponibili fanno ritenere che essi svolgano un ruolo centrale nella conservazione dell'omeostasi. Gli eventi associati al LES rappresentano una disfunzione dei meccanismi omeostatici normali, disfunzione che si manifesta clinicamente con una vasta gamma di sintomi.

A metà del diciannovesimo secolo il lupus eritematoso veniva considerato semplicemente una malattia della pelle, ma alla fine del secolo Moritz Kaposi a Vienna e William Osler della Johns Hopkins University descrissero una serie di manifestazioni sistemiche spesso associate al lupus facciale: artrite, polmoniti e un certo numero di sintomi riguardanti il sistema nervoso centrale. Nella prima parte di questo secolo al quadro clinico di questa malattia vennero aggiunte patologie di parecchi organi interni. Nel 1941 il LES fu classificato da Paul Klemperer, Abou D. Pollack e George Baehr del Mount Sinai Hospital di New York come una «malattia collageno-vascolare», una malattia cioè che colpisce il tessuto connettivo e i vasi sanguigni.

Fu la scoperta delle cellule del lupus eritematoso (cellula LE) nel 1948 da parte di M. M. Hargraves della Mayo Clinic che richiamò l'attenzione sugli aspetti immunologici del LES. Halsted R. Holman e Kunkel trovarono che il siero dei pazienti affetti da LES conteneva una gammaglobulina diretta contro i nuclei e specificamente contro le nucleoproteine, le proteine che formano un complesso con il DNA nei cromosomi. Da queste ricerche risultò che nel siero dei pazienti sono presenti svariati anticorpi che interagiscono con i nuclei, e si dimostrò che uno di questi era responsabile del fenomeno delle cellule LE. L'osservazione che la maggior parte dei sieri dei pazienti affetti da LES contiene anticorpi reattivi verso gli antigeni dello stesso organismo suffragò l'opinione che il LES fosse una malattia autoimmune.

Tuttavia è difficile spiegare i sintomi tipici dei pazienti affetti da LES per mezzo delle reazioni autoimmuni. Tra questi sintomi vi sono febbre, spossatezza, eruzioni cutanee, artrite, infiammazione vascolare, pleurite, pericardite e glomerulonefrite, la malattia renale caratterizzata dall'infiammazione dei glomeruli del rene, che è una delle manifestazioni più pericolose del LES. Gli anticorpi antinucleari, che sono piuttosto comuni, di solito non penetrano attraverso le membrane delle cellule viventi intatte; inoltre nei tessuti dei pazienti affetti da LES non si ritrovano depositi di anticorpi antinucleari, come invece ci si aspetterebbe se essi fossero la causa diretta della patologia.

Le manifestazioni cliniche e patologiche del LES assomigliano molto più strettamente a quelle della malattia da siero che a quelle di una malattia autoimmune tipica ed è noto che la malattia da siero è provocata da immunocomplessi. Le prove che suggeriscono un ruolo patogeno diretto da parte degli immunocomplessi derivano da parecchi tipi di ricerche diverse. Per prima cosa, pressappoco nello stesso periodo in cui si scoprirono gli anticorpi diretti contro il nucleo nel siero dei pazienti affetti da LES, Robert C. Mellors del Cornell University Medical College trovò depositi di gammaglobuline nei glomeruli renali dei pazienti. Jacinto J. Vazquez e Frank J. Dixon, che allora si trovavano alla University of Pittsburgh School of Medicine, trovarono depositi simili nelle piccole arterie del fegato e della milza e nelle cellule LE. Si trovò che gli anticorpi circolanti erano diretti principalmente contro gli antigeni nucleari: il DNA, sia nella forma naturale a doppio filamento che nella forma denaturata a filamento singolo, e i complessi di DNA e istoni. Eng M. Tan, Kunkel e i loro colleghi dimostrarono che nel sangue di parecchi pazienti il livello del DNA con funzioni di antigene e quello degli anticorpi anti-DNA mostrano ampie fluttuazioni, suggerendo una correlazione tra la formazione di immunocomplessi e le fasi di riacutizzazione della malattia. Studi successivi hanno confermato un'impressionante associazione tra le fasi acute ricorrenti della malattia e una riduzione della concentrazione nel plasma del complemento, che si lega ai complessi antigene-anticorpo e si deposita nei tessuti o viene rimosso dal sistema reticoloendoteliale.

Nel grafico sono rappresentati gli eventi connessi con un peggioramento della malattia in un paziente affetto da LES con limitate manifestazioni renali. Il contenuto di anticorpi ematici contro il DNA a filamento singolo e doppio aumenta e la concentrazione di complemento incomincia a diminuire, finché compare la proteinuria (un aumento dell'escrezione di proteine nelle urine), un segnale di disfunzione renale. Si noti la rapida diminuzione degli anticorpi associata all'estrema depressione del livello di complemento; gli anticorpi si sono combinati con l'antigene a formare gli immunocomplessi sui quali si fissa il complemento (e che possono essere osservati nel siero sotto forma di crioglobulina, un precipitato gelatinoso che si forma a bassa temperatura). Un altro anticorpo antinucleo, l'anti-Sm, non mostra un apprezzabile aumento durante la crisi; sembra evidente che solo alcuni anticorpi partecipano alla formazione degli immunocomplessi capaci di produrre danni ai tessuti. Il prednisone, un farmaco steroideo antiinfiammatorio, è efficace nel trattamento degli episodi acuti anche se non si conosce ancora il suo meccanismo di azione.

Infine, insieme a Kunkel e ad altri suoi collaboratori riuscii a dimostrare visivamente che i complessi antigene-anticorpo possono costituire importanti mediatori delle lesioni ai vasi sanguigni e ai reni nel LES, a seconda del luogo di deposizione. Esaminammo campioni di biopsie renali prelevate ai pazienti in varie fasi della malattia. L'osservazione al microscopio ottico con tecniche di colorazione convenzionali rivelò danni ai tessuti di diversa entità. L'osservazione con il microscopio a fluorescenza di diversi campioni esposti a un anticorpo marcato con un colorante fluorescente rivelò che la maggior parte dei pazienti affetti da LES mostrava depositi di immunocomplessi nel mesangio, la regione del glomerulo che riassorbe le molecole di maggiori dimensioni.

Quando questi depositi erano limitati al mesangio, i pazienti mostravano poca o nessuna proteinuria (presenza di proteine in eccesso nelle urine, il principale segnale per la diagnosi di una glomerulonefrite). Nei campioni prelevati dai pazienti con un aumento di proteinuria, d'altra parte, depositi granulari di gammaglobuline e complemento rivelarono la presenza di complessi a livello della membrana basale, dove il sangue viene filtrato. La prolungata deposizione di questi complessi dà origine a un ispessimento delle pareti dei capillari. Proprio l'osservazione di questi ispessimenti nei glomeruli dei pazienti affetti da LES aveva indotto Klemperer nel 1935 a stabilire una correlazione tra LES e lesioni della malattia da siero prodotte nei reni di cavalli a cui erano state praticate ripetutamente iniezioni di tossina difterica per preparare l'antisiero.

Il ruolo dei complessi nel LES rimase però ipotetico finché Frederick G. Germuth, Jr., che allora si trovava alla Johns Hopkins, nel 1950 eseguì esperimenti su animali stabilendo definitivamente l'importanza dei complessi e correlando specificamente la gravità della malattia con la risposta immunitaria verso un antigene. Germuth e i suoi còlleghi dimostrarono che la somministrazione prolungata di antigeni proteici ai conigli dava origine a una malattia cronica simile alla malattia da siero degli animali e a lesioni glomerulari simili a quelle riscontrate nei pazienti affetti da LES. Se i conigli venivano immunizzati con una dose costante di una proteina (la sieroalbumina bovina), la gravità della glomerulonefrite che ne seguiva dipendeva dalla risposta anticorpa-

le del singolo coniglio. Nei conigli che rispondevano intensamente (con massicce dosi di anticorpo) si formavano grandi aggregati paracristallini di complessi antigene-anticorpo; questi venivano rapidamente rimossi dal sistema reticoloendoteliale e non sembravano essere tossici. Nei conigli che rispondevano meno intensamente si formavano complessi leggermente più piccoli, che tendevano a concentrarsi nel mesangio. Nei conigli con una risposta anticorpale ancora meno intensa si formavano complessi ancora più piccoli; questi conigli venivano colpiti da glomerulonefrite diffusa proliferativa, la forma più grave della malattia. Risposte anticorpali ancora più deboli davano origine a complessi ancora più piccoli e i conigli non mostravano segni di malattia.

Studi precedenti di Theodore P. Pincus e Charles L. Christian del Columbia University College of Physicians and Surgeons avevano indicato che la tossicità dei complessi era influenzata dalle proprietà intrinseche degli anticorpi contro la sieroalbumina bovina. Germuth, sulla base di ricerche successive condotte alla Saint Louis University School of Medicine, sottolineò il ruolo di una caratteristica essenziale dell'anticorpo, la sua forte tendenza a legarsi con l'antigene. Dixon e i suoi colleghi della Scripps Clinic eseguirono una serie di esperimenti di fondamentale importanza; essi fecero in modo di somministrare dosi di antigene in leggero eccesso rispetto all'anticorpo corrispondente. I complessi così formati dimostrarono una particolare tendenza a provocare la glomerulonefrite membranosa cronica. L'instaurarsi della nefrite era associato a un significativo aumento di depositi renali e ciò faceva ritenere che la lesione iniziale favorisse l'aumento dei depositi; un possibile fattore facilitante poteva essere costituito dalla saturazione del sistema reticoloendoteliale, che in tal modo non era più in grado di eliminare dal sangue i complessi antigene-anticorpo. Robert T. McCluskey e Baruj Benacerraf, che allora si trovavano alla New York University School of Medicine, furono in grado di provocare glomerulonefrite e infiammazione dei vasi sanguigni nei topi somministrando direttamente immunocomplessi solubili che essi avevano preparato in laboratorio.

In sintesi, il potenziale tossico degli immunocomplessi è variabile e questo spiega in parte l'ampiezza dello spettro di sintomi clinici. Il luogo di deposizione dei complessi e lo scatenarsi di una risposta infiammatoria sono influenzati dalla dimensione degli aggregati, dal numero di siti di legame di antigeni e anticorpi, dal numero di anticorpi diversi presenti e dalla loro affinità per l'antigene. Charles G. Cochrane della Scripps Clinic trovò inoltre che anche le sostanze che incrementavano la permeabilità dei vasi sanguigni aumentano la tossicità dei complessi. Trattando i conigli con antagonisti dell'istamina e della serotonina si riduceva il deposito di complessi e la conseguente infiammazione. Mart Mannik della University of Washington School of Medicine dimostrò che la capacità del sistema reticoloendoteliale di rimuovere i complessi è condizionata dalla capacità dei fagociti di riconoscere una particolare regione dell'anticorpo.

La glomerulonefrite, un'infiammazione delle anse capillari dei glomeruli del rene, è una grave manifestazione del lupus eritematoso sistemico. Il rene *(a)*, dove i fluidi e le sostanze tossiche del sangue vengono rimossi ed escreti nelle urine, è composto da circa un milione di unità funzionali chiamate nefroni. Ogni nefrone *(b)* consiste di un groviglio di capillari piegati ad ansa, il glomerulo, dove il sangue è filtrato, e di un tubulo attraverso il quale il filtrato passa a un dotto collettore; gran parte del fluido e alcune delle sostanze in soluzione sono riassorbite dal tubulo in una rete di capillari che circonda il tubulo. Il glomerulo *(c)* è costituito da un fitto intreccio di capillari. Una sezione del glomerulo *(d)* mostra il lume interno delle diramazioni dei capillari avvolti in una membrana basale ricoperta di cellule epiteliali; la membrana e le cellule epiteliali sono deputate al controllo della filtrazione. Le grosse molecole vengono rimosse dal sangue e depositate nel mesangio.

In alcuni pazienti affetti da LES con sintomi a carico dei reni, la malattia rimane limitata a una glomerulonefrite a focolai, in cui solo alcuni glomeruli presentano depositi di immunocomplessi e sono danneggiati; i farmaci che sopprimono la risposta immunitaria o riducono l'infiammazione sembrano essere particolarmente efficaci in questo stadio della malattia. In altri pazienti la malattia avanza fino alla fase diffusa e proliferativa caratterizzata da massicci depositi nella maggior parte dei glomeruli e da una significativa proteinuria. Recentemente Giuseppe A. Andres della State University of New York School of Medicine a Buffalo e McCluskey, ora alla Harvard Medical School, hanno sottolineato l'importanza delle lesioni ai tubuli renali (che riassorbono i fluidi e le sostanze in soluzione dalla massa di liquido filtrata dai glomeruli) come effetto della deposizione dei complessi nei tessuti che circondano il tubulo.

Con lo scopo di chiarire la natura dei complessi che si trovano nelle forme progressive della malattia, Kunkel, Peter H. Schur e io esaminammo i reni di pazienti morti a causa dei danni renali provocati dal LES; John B. Winfield e io analizzammo le crioglobuline (immunocomplessi precipitati dal siero per effetto delle basse temperature) di pazienti affetti da LES. Riuscimmo a ottenere mediante eluizione parecchi anticorpi glomerulari e a isolare anticorpi simili dalle crioglobuline. Trovammo che gli anticorpi ottenuti con questi due metodi hanno concentrazione più elevata di quella riscontrata nel siero non frazionato, il che si spiega supponendo che vengano concentrati nei complessi antigene-anticorpo. Questi anticorpi erano diretti contro il DNA a doppio filamento, contro il DNA a filamento singolo e contro le ribonucleoproteine; i complessi contenenti DNA erano di gran lunga i più frequenti, ed è stato possibile dimostrare la presenza di DNA con funzione di antigene nei glomeruli contenenti gli anticorpi anti-DNA.

L'immunocomplesso contenente DNA è stato il primo agente specifico in grado di danneggiare i tessuti identificato nella glomerulonefrite dell'uomo; altri antigeni sono stati localizzati nei reni di pazienti affetti da glomerulonefrite «primaria», una malattia renale simile a quella provocata dal LES, ma non associata a questo. Poiché i molti sintomi del LES in fase iniziale segnalano il coinvolgimento della funzione renale, questa malattia è stata studiata come prototipo delle due malattie causate dagli immunocomplessi.

Winfield, Kunkel e io abbiamo trovato che gli anticorpi anti-DNA eluiti dai glomeruli hanno un'acidità (termine usato da alcuni autori in sostituzione di avidità perché più generico) più alta verso il loro antigene rispetto allo stesso tipo di anticorpi presenti nel siero di individui sani. Inoltre, gli anticorpi anti-DNA del siero di pazienti affetti da malattia renale hanno un'avidità più bassa di quelli dei controlli. Queste scoperte fanno ritenere che nella glomerulonefrite gli anticorpi sopra ricordati vengano depositati preferenzialmente sotto forma di complessi nei glomeruli renali. Il risultato è che la semplice analisi del siero per riconoscere gli anticorpi circolanti può non essere appropriata per rivelare la presenza di un particolare anticorpo perché questo è legato in un complesso. Sono stati fatti notevoli progressi nei metodi per identificare i complessi circolanti e valutare così la gravità della malattia; è ora possibile per esempio isolare i componenti del complemento e comprenderne i meccanismi di reazione.

Per individuare i complessi è possibile esaminare le crioglobuline, le cui interazioni con il complemento sono state studiate da Christian. Vincent Agnello, Kunkel e io abbiamo trovato che i pazienti con quantità particolarmente ingenti di crioglobuline presentano massicci depositi di IgM nei reni, a differenza dei pazienti affetti dalla malattia renale associata al LES, in cui l'immunoglobulina principale è l'IgG; il fattore reumatoide IgM è anche uno dei componenti principali delle crioglobuline circolanti del LES. Queste scoperte indicano che i complessi di gammaglobuline e dei loro anticorpi contribuiscono al danno renale in almeno un gruppo di pazienti. Recentemente Morris Reichlin della State University of New York School of Medicine a Buffalo ha scoperto un altro immunocomplesso che coinvolge l'attività di antigeni provenienti dal citoplasma delle cellule, avvalorando l'ipotesi che complessi diversi possono contribuire al danno renale.

Agnello, Robert J. Winchester e Kunkel osservarono che il C1q (la proteina del complemento che interagisce con gli anticorpi) forma un precipitato con gli aggregati di immunoglobuline e questo fornì un metodo per svelare la presenza di complessi solubili nel plasma. Trovammo, infatti, questi precipitati nel siero dei pazienti affetti da LES. Paul H. Lambert e Peter A. Miescher dell'Università di Ginevra hanno messo a punto un altro metodo per il dosaggio dei complessi che utilizza la proteina C1q marcata con un isotopo radioattivo; mentre Argyrios N. Theofilophoulos e Dixon hanno messo a punto un sistema che sfrutta alcune cellule linfoidi (chiamate cellule di Rajii) che possiedono recettori per i componenti del complemento fissati ai complessi. Queste e altre metodiche per il riconoscimento dei complessi antigene-anticorpo forniscono informazioni sul decorso del LES che non potrebbero essere ottenute con un semplice controllo del livello di antigene o di anticorpo.

Mentre è stata dimostrata una chiara relazione tra gli immunocomplessi da una parte e la malattia renale e le lesioni ai vasi sanguigni dall'altra, è stato difficile identificare i meccanismi con cui vengono prodotti danni ad altri tessuti in alcune manifestazioni del LES. Indagini compiute per mezzo della tecnica dell'immunofluorescenza su biopsie di pelle hanno dimostrato per esempio che in circa metà dei pazienti affetti da LES si riscontrano dei depositi di immunocomplessi appena sotto lo strato più esterno della pelle (l'epidermide). Tuttavia non sempre questi depositi sono causa di infiammazione. Tan e Kunkel hanno osservato depositi granulari di gammaglobuline sia nelle lesioni infiammatorie sia in zone di pelle assolutamente normali. Morris Ziff e James N. Gilliam dell'University of Texas Health Sciences Center di Dallas e Reichlin hanno riferito che l'incidenza di questi depositi nelle zone di pelle normale è più alta nei pazienti con lesioni renali, i quali spesso hanno nel sangue una maggiore quantità di immunocomplessi. I depositi possono venir riconosciuti anche in quelle forme di lupus in cui le alterazioni della pelle costituiscono il sintomo principale. Molte domande su queste lesioni e sui depositi a esse associati attendono ancora risposta. Che cosa differenzia l'effetto dei complessi nella pelle normale da quello nella pelle infiammata: le proprietà intrinseche dei depositi o la risposta del paziente? I complessi depositati nella pelle sono uguali o differenti da quelli depositati nei glomeruli? Se sono diversi, qual è il motivo di questa differenza?

La base dei sintomi neurologici del LES, che generalmente sono rilevabili solo alcuni anni dopo l'inizio della malattia e che ora sono più frequenti a causa dell'aumento della vita media dei pazienti colpiti dal LES, è ancora oscura. Non ci sono indizi di un collegamento tra depositi e lesioni al cervello e neppure è provato il collegamento di particolari lesioni con sintomi definiti come convulsioni e psicosi. Nel siero dei pazienti affetti da LES è stata trovata una notevole varietà di anticorpi che reagiscono con antigeni del sistema nervoso, ma questi non sembrano coinvolti in disturbi neurologici.

Tutto ciò che possiamo dire è che esiste un effetto diretto degli autoanticorpi (da distinguere dai complessi) presenti nei pazienti affetti da LES sugli elementi corpuscolari del sangue; gli anticorpi che reagiscono con gli antigeni presenti sulla superficie dei globuli rossi possono essere associati all'anemia emolitica, una condizione patologica in cui i globuli rossi sono danneggiati e il livello di emoglobina è di conseguenza ridotto.

Nel siero di pazienti affetti da LES si

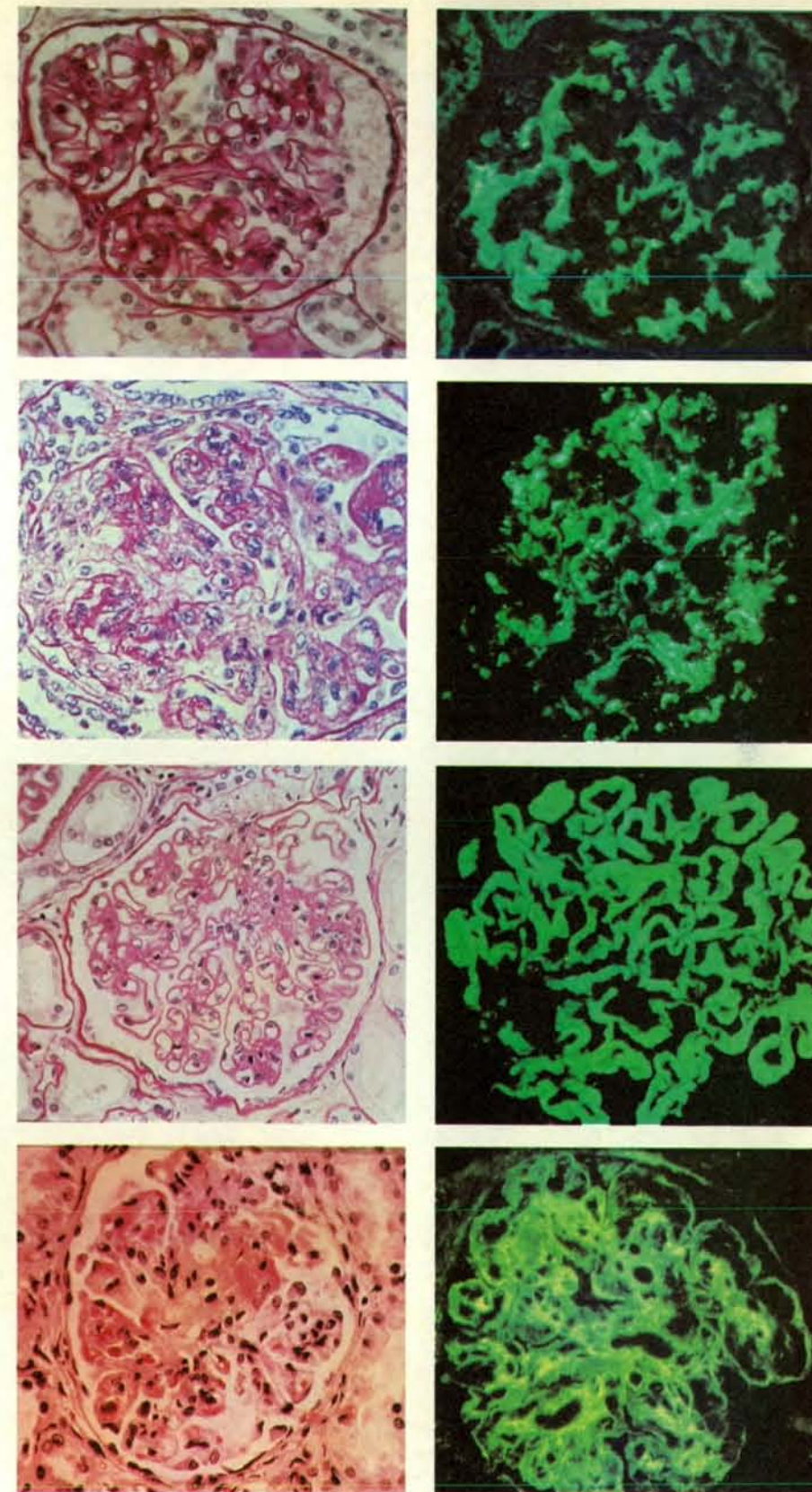
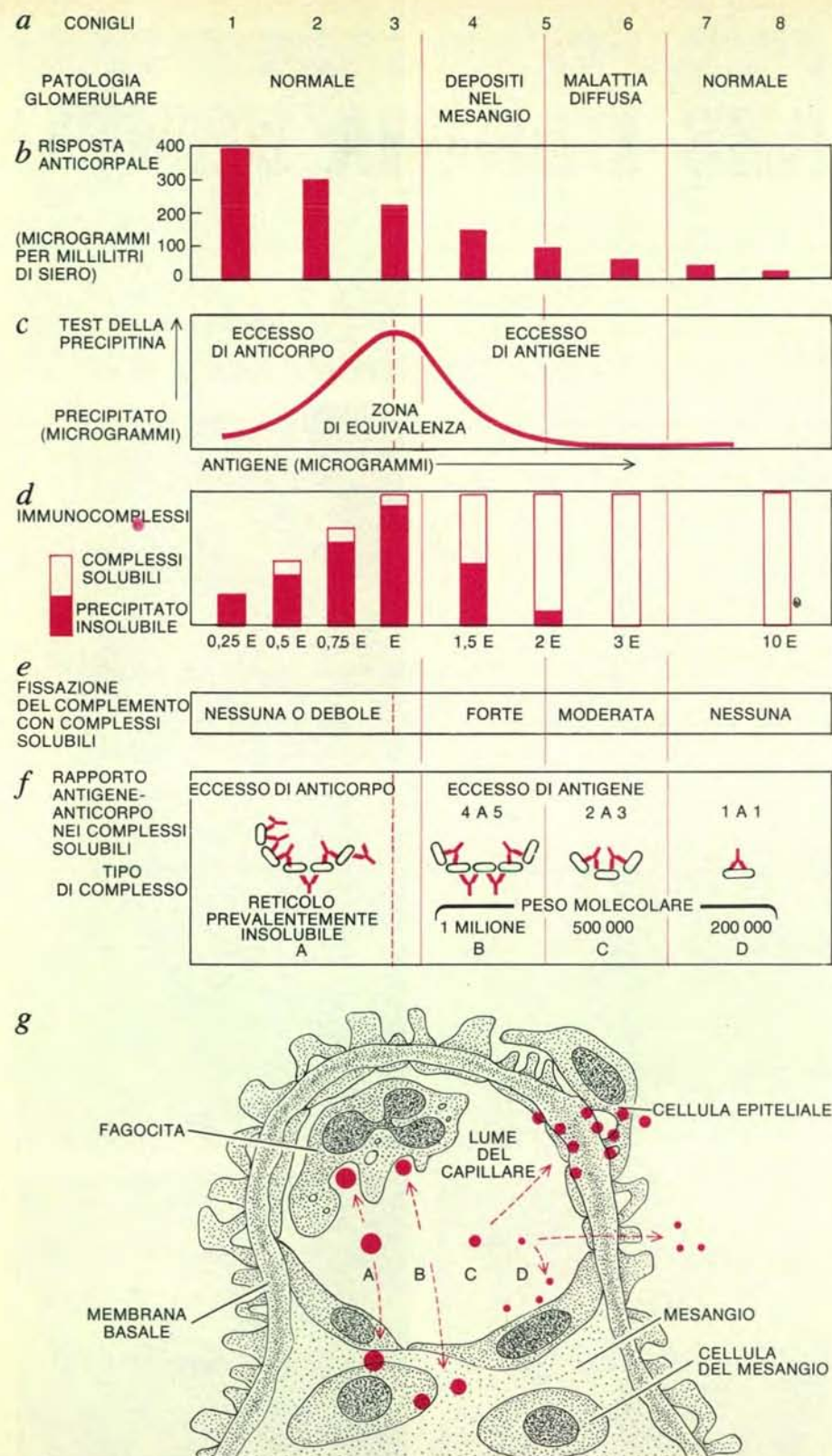
trovano anticorpi contro i linfociti di tipo *B* e *T*. Sono state osservate disfunzioni delle cellule *T*, compresa una riduzione dell'attività soppressiva verso le cellule *B*; una popolazione di anticorpi contro le cellule *T* con antigeni simili a quelli del cervello di ratto, ma non è ancora noto se questi anticorpi alterino il funzionamento dell'immunità cellulare o interagiscano con i tessuti cerebrali dei pazienti. Winfield, Winchester e Kunkel hanno notato che la presenza di alcuni anticorpi antilinfociti nel siero dei pazienti affetti da LES è correlata a una riduzione del livello di linfociti circolanti. Essi hanno anche trovato un'alta concentrazione di questi anticorpi nelle crioglobuline: questo può voler dire che essi possono formare immunocomplessi con gli antigeni di superficie staccatisi dai linfociti in circolo.

A parte questi anticorpi responsabili dell'anemia emolitica, tuttavia, non è stata chiaramente dimostrata la responsabilità degli autoanticorpi per le lesioni dei vari tessuti nel LES. Vi è ora consenso generale sul fatto che sono gli immunocomplessi a causare le lesioni dei vasi sanguigni e dei glomeruli. Altri studi di Andres sulla membrana sinoviale delle articolazioni, sulla pleura e sul pericardio (le membrane che avvolgono rispettivamente i polmoni e il cuore) e sulle pareti degli alveoli polmonari fanno ritenere che i complessi possano essere la causa di molti dei sintomi del LES.

In questi disegni è schematizzato il meccanismo di deposizione nel rene di immunocomplessi durante i differenti stadi del lupus sistemico. Anche in un paziente non affetto dalla malattia renale il glomerulo strutturalmente normale presenta depositi di immunocomplessi nel mesangio *(in alto)*. Due terzi dei pazienti affetti da LES vengono colpiti dalla malattia renale; di questi il 70 per cento presenta una glomerulonefrite proliferativa a focolai con proteinuria minima o assente *(al centro)* e depositi puntiformi lungo la membrana basale sotto il rivestimento di cellule endoteliali. Il 15 per cento circa viene colpito da glomerulonefrite membranosa, con proteinuria moderata o marcata *(in basso a sinistra)*; la prolungata deposizione di complessi produce depositi piuttosto omogenei che delineano la membrana basale. Il restante 15 per cento (e alcuni dei pazienti con glomerulonefrite a focolai) viene successivamente colpito da glomerulonefrite diffusa proliferativa, con massicci depositi di complessi nella maggior parte dei glomeruli *(in basso a destra)*. Questo stadio conduce di solito a una insufficienza renale.

Finora ho discusso i meccanismi con cui il LES provoca danni ai tessuti. Dire che il LES sembra una malattia dovuta agli immunocomplessi non serve tuttavia a individuarne le cause, di cui per ora si sa molto poco. È noto, per esempio, che esiste una predisposizione genetica al LES, che nella sindrome del LES sono implicite delle anomalie del sistema immunitario del paziente e che nella malattia possono essere coinvolti svariati fattori come certi farmaci, gli ormoni estrogeni, le radiazioni ultraviolette e alcune infezioni virali.

La prova più evidente di una predisposizione genetica al LES è il fatto che se uno di due gemelli omozigoti viene colpito dal LES, l'altro gemello viene colpito nel 50 o 60 per cento dei casi, mentre il tasso di concordanza è solo del 2 o 3 per cento nel caso dei gemelli eterozigoti e di altri parenti di primo grado dei pazienti affetti da LES; l'incidenza nel complesso della popolazione è appena dello 0,2 per cento circa. Al LES sono state associate parecchie anomalie congenite delle proteine del complemento. I geni che controllano la sintesi di alcune delle proteine del complemento si trovano sullo stesso cromosoma dei geni che controllano la risposta immunitaria e la produzione degli antigeni di istocompatibilità, un gruppo di molecole caratteristico per ogni individuo il quale determina il rigetto dei trapianti. L'incidenza del LES è molto più alta per gli individui che portano determinati antigeni di istocompatibilità. Tutto ciò dimostra che esiste una componente genetica nel LES, ma i meccanismi per cui i fattori genetici sono coinvolti nel LES devono ancora essere chiariti. Il collegamento tra un fattore genetico e la malattia può dipendere da qualche fattore ausiliario. Per esempio, un difetto del sistema complemento o del sistema immunitario può costituire una predisposizione a un'infezione virale che a sua volta può essere la causa diretta della malattia.

Nel LES sono state dimostrate disfunzioni sia della componente cellulare che di quella anticorpale del sistema immunitario. I pazienti posseggono molti differenti autoanticorpi con una particolare propensione a formare immunocomplessi e uno di questi, l'anticorpo contro il DNA a doppio filamento, è presente quasi solo nel LES. Una sottopopolazione di anticorpi anti-DNA capaci di formare complessi che legano il complemento è stata strettamente correlata con la distruzione dei tessuti. La rarità di questi anticorpi in

Questi campioni di tessuto prelevati a quattro diversi stadi della malattia mostrano i cambiamenti strutturali evidenziati dalla colorazione istologica *(a sinistra)* e la localizzazione degli immunocomplessi per mezzo della microscopia a fluorescenza *(a destra)*. Il grumerulo normale *(in alto a sinistra)* presenta depositi irregolari di immunocomplessi nelle zone corrispondenti al mesangio *(in alto a destra)*. Un gromerulo di un paziente con glomerulonefrite a focolai *(secondo dall'alto)* mostra un ispessimento irregolare della membrana basale e la fusione di capillari; nelle pareti dei capillari sono immobilizzati depositi granulari di complessi. Nella glomerulonefrite membranosa *(terza dall'alto)* vi è un ispessimento più marcato della membrana *(a sinistra)* e i depositi di immunocomplessi assumono l'aspetto della microfotografia a destra. Nella glomerulonefrite diffusa *(in basso)* profondi danni dei tessuti *(a sinistra)* e un massiccio accumulo di immunocomplessi *(a destra)* alterano la normale struttura del glomerulo e bloccano la filtrazione.

Frederick G. Germuth, Jr., della Johns Hopkins University School of Medicine ha confrontato le dimensioni dei complessi con la risposta anticorpale e la gravità della malattia nei conigli. In seguito alla somministrazione regolare e costante di antigene, la patologia riscontrata in otto conigli *(a)* era correlata alla quantità di anticorpo formato da ciascun animale *(b)*. Un test di precipitazione (eseguito *in vitro* aggiungendo quantità crescenti di antigene a una quantità costante di anticorpo) dimostra che la quantità di immunocomplessi precipitati (insolubili) varia in base al rapporto anticorpo-antigene *(c)*. La curva viene così interpretata *(d)*: la quantità di precipitato aumenta finché l'antigene è in quantità equivalente all'anticorpo *(E)*, poi decresce. I complessi differiscono tra di loro per la capacità di legare il complemento, che determina la tossicità *(e)*. Le dimensioni del complesso, correlate al rapporto antigene-anticorpo *(f)*, determinano il sito di deposizione *(g)*. I complessi molto grandi *(A, B)* vengono trasferiti entro le cellule del mesangio e i fagociti in circolo; i complessi di peso molecolare intorno a 500 000 *(C)* vengono depositati lungo la membrana del glomerulo e provocano la malattia renale; i complessi molto piccoli o rimangono in circolo o vengono filtrati nelle urine senza alcun danno al tessuto renale.

altre malattie e il fatto che non si possa indurre la formazione di anticorpi immunizzando gli animali con il DNA a doppio filamento fanno ritenere che il sistema immunitario sia normalmente refrattario a riconoscere il DNA a doppio filamento come un antigene estraneo. Ciononostante sono state trovate cellule *B* con recettori per il DNA a doppio filamento nel sangue di individui normali, e Arthur D. Bankhurst e Ralph C. Williams, Jr., della University of New Mexico School of Medicine ne hanno trovato in numero ancora maggiore in pazienti affetti da LES. Sembra che cloni di cellule capaci di produrre anticorpi contro questo particolare autoantigene si siano in qualche modo sottratti, nei pazienti affetti da LES, al complesso sistema di pesi e contrappesi che normalmente regola l'attività del sistema immunitario.

Come può avvenire una cosa del genere? Molte possibili spiegazioni sono suggerite dallo studio dei meccanismi dell'immunità cellulare. Sia i pazienti affetti da LES sia un ceppo di topi che presentano una serie di sintomi simili a quelli del LES hanno una ridotta attività di quelle cellule *T* soppressive. L'attività di queste cellule decresce con l'età nei topi con una sindrome simile al LES; questa diminuzione è associata a un aumento della produzione di anticorpi anti-DNA e con una forma virulenta di glomerulonefrite. Alfred D. Steinberg, Anthony S. Fauci, M. Eric Gershwin e Thomas A. Waldmann dei National Institutes of Health hanno presentato prove convincenti di un funzionamento difettoso delle cellule *T* soppressive e hanno ottenuto informazioni sul funzionamento delle cellule *B* sia nei topi sia nei pazienti colpiti da LES.

Nell'uomo è stato dimostrato che il livello di anticorpi antinucleari aumenta con l'età, forse a causa di una riduzione dell'attività delle cellule soppressive; a questo aumento non è però corrisposto alcun sintomo. È stato suggerito che alcuni anticorpi diretti contro le cellule *T* riscontrabili sia nell'uomo sia nel topo contribuiscono alle anomalie del funzionamento delle cellule soppressive. Peter Wernet e Kunkel hanno dimostrato l'esistenza di anticorpi che interagiscono con le cellule *T* e ne inibiscono la proliferazione. Tale inibizione, nel caso che si verificasse anche *in vivo*, tenderebbe a svincolare le cellule *B* dalla soppressione esercitata dalle cellule *T*. Robert S. Schwartz della Tufts University School of Medicine ha recentemente rilevato delle anomalie a carico delle cellule soppressive nelle famiglie dei pazienti affetti da LES. D'altra parte è possibile che qualche fattore come l'interazione con gli immunocomplessi presenti in circolo possa provocare la proliferazione delle cellule *T helper* e di conseguenza stimolare la produzione di anticorpi da parte dei linfociti *B*.

I risultati di una indagine a lungo termine condotta su cani colpiti da LES da Schwartz e Robert M. Lewis suggerisco-

a ANALISI DEGLI ANTICORPI

TESSUTO

CORTICALE DEL RENE

OMOGENEIZZAZIONE E CENTRIFUGAZIONE

GLOMERULI

GLOMERULI TRATTATI CON TAMPONE ACIDO O DNASI PER ELUIRE GLI ANTICORPI

ANTICORPI

SIERO

SIERO DEL PAZIENTE A 37 °C, CRIOGLOBULINE IN SOLUZIONE

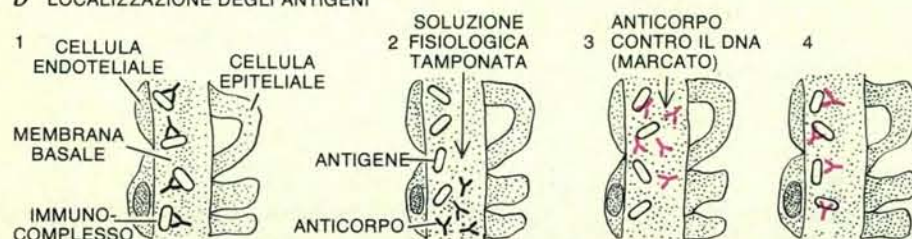
SIERO INCUBATO A 4 °C, CRIOGLOBULINE PRECIPITATE

PRECIPITATO DI CRIOGLOBULINE

PRECIPITATO RIDISCIOLTO, INCUBATO A 4 °C E RIPRECIPITATO

CRIOGLOBULINE

ANTICORPI CONTRO:	ELUITO GLOMERULARE	CRIOGLOBULINE
DNA A FILAMENTO DOPPIO	6 SU 9	21 SU 29
DNA A FILAMENTO SINGOLO	8 SU 9	16 SU 28
RIBONUCLEOPROTEINE NUCLEARI	3 SU 9	3 SU 13

b LOCALIZZAZIONE DEGLI ANTIGENI

La natura dei complessi associati alla malattia renale grave è stata stabilita *(a)* esaminando, in campioni ottenuti da autopsie, i glomeruli e le crioglobuline, una frazione del siero che contiene principalmente immunocomplessi e forma precipitati gelatinosi a bassa temperatura. Si è trovato che certi anticorpi sono presenti in concentrazione molto più alta nei glomeruli e nelle crioglobuline che nel siero, indicazione di una forte tendenza a formare complessi. La tabella mostra, per ognuno dei tre anticorpi, la proporzione di casi in cui questa concentrazione preferenziale era di almeno 50 volte. Ulteriori prove della presenza degli immunocomplessi nei glomeruli è stata ottenuta *(b)* dissociando gli anticorpi dagli antigeni in sezioni di rene, aggiungendo anticorpo anti-DNA marcato e riscontrando il suo legame al DNA antigenico negli stessi siti di deposito dei complessi.

no la partecipazione di fattori sia ambientali che genetici. Questi cani presentano una sindrome notevolmente simile al LES umano, compresi l'eczema caratteristico e la presenza di anticorpi specifici nel siero. Tuttavia dopo un prolungato incrocio (selezione di linee pure) alcuni cani mostrano le anomalie del siero senza alcun sintomo di malattia, come spesso succede per i parenti dei pazienti umani colpiti da LES. Lawrence E. Schulman della Johns Hopkins ha concluso dalla sua analisi condotta sulle famiglie dei pazienti affetti da LES che pare dimostrata l'esistenza di fattori sia genetici che ambientali.

Se i presupposti per l'insorgenza del LES clinicamente rilevabile sono sia una predisposizione genetica sia l'intervento di un fattore precipitante, quale potrebbe essere la natura di un tale fattore? Molti farmaci sono stati associati a un aumento dell'incidenza di anticorpi antinucleari. In particolare la procainammide (prescritta per la cura delle aritmie cardiache) e la idralazina (contro l'ipertensione) inducono la produzione di anticorpi contro il DNA a singolo filamento, contro le nucleoproteine e forse contro altri antigeni nucleari; la somministrazione prolungata di procainammide dà origine a una sindrome simile al LES in circa un terzo dei soggetti trattati. I sintomi indotti da questo farmaco scompaiono tuttavia con la sospensione del trattamento e raramente includono alterazioni dei reni o produzione di anticorpi contro il DNA a doppio filamento.

Come modello sperimentale, tuttavia, il LES indotto da farmaci offre la possibilità di studiare il meccanismo di induzione della produzione di autoanticorpi. Per esempio, la procainammide e l'idralazina vengono metabolizzate mediante un'acetilazione, e Marcus M. Reidenberg e Dennis E. Drayer della Cornell University hanno trovato che gli effetti del farmaco sono più pronunciati negli «acetilatori lenti» che sono individui con un basso livello, geneticamente programmato, di enzima acetilante. Robert G. Lahita, Reidenberg, Drayer e io abbiamo dimostrato che i pazienti a cui viene somministrato un analogo acetilato della procainammide invece che la forma usuale del farmaco producono gli autoanticorpi più lentamente e in quantità minore. Finora non ci sono prove che indichino un farmaco o un'altra sostanza chimica come un fattore precipitante del LES di per sé, ma questa possibilità deve essere ancora verificata.

Parecchie osservazioni suggeriscono un'associazione con gli ormoni estrogeni. Il LES colpisce prevalentemente le donne durante gli anni di maturità sessuale, quando la concentrazione di estrogeni è massima; l'uso di contraccettivi orali contenenti estrogeni talvolta provoca l'aggravamento della malattia; gli uomini sono colpiti dal LES solo più raramente, ma la malattia è più frequente nella rara condizione congenita chiamata sindrome di Klinefelter, in cui l'individuo di sesso maschile possiede un cromosoma *Y* e due cromosomi *X* (invece che uno solo) e ha un metabolismo anomalo degli estrogeni. Norman Talal della University of California School of Medicine a San Francisco e i suoi colleghi hanno dimostrato che è possibile ridurre l'intensità del LES dei topi rimuovendo le gonadi delle femmine e privandole così degli estrogeni, oppure somministrando loro ormoni maschili, e che la malattia viene aggravata castrando i topi maschi o somministrando loro estrogeni. Recenti studi di Lahita, Kunkel e Jack Fishman hanno dimostrato che nei pazienti affetti da LES, sia maschi che femmine, si trovano livelli anormalmente alti di metaboliti degli estrogeni.

La sensibilità alla luce solare dimostrata generalmente dai pazienti affetti da LES implica che la radiazione ultravioletta possa essere un fattore precipitante. Le lesioni cutanee tendono a comparire in regioni del corpo esposte alla luce solare; occasionalmente vengono aggravati anche i sintomi sistemici. Le radiazioni ultraviolette possono alterare i filamenti del DNA a doppio filamento. Alcune delle strutture anomale risultanti, come i dimeri di timidina (in cui due basi nucleotidiche del DNA sono legate covalentemente), possono essere utilizzate come epitopi immunogeni negli animali di laboratorio. Le radiazioni ultraviolette potrebbero potenziare la risposta immunitaria e abolire la tolleranza naturale al DNA.

Il ruolo delle infezioni virali nel LES è stato argomento di molte ipotesi e ricerche, la maggior parte delle quali si sono incentrate sui virus a RNA del tipo *C*. Nei tessuti colpiti dal LES sono state ripetutamente osservate proteine virali, ma non è chiaro quanto questo sia significativo. L'infezione virale potrebbe non essere un fattore precipitante, ma piuttosto una conseguenza di una deficienza immunitaria o della soppressione della risposta immunitaria dovuta ai farmaci prescritti per la cura del LES; il materiale genetico virale potrebbe trovarsi all'interno dei tessuti umani senza produrre in effetti altre particelle virali e senza avere nulla a che fare con il decorso del LES. Prove circostanziate di una causa virale del LES sono state presentate da Raphael J. De Horatius e Ronald P. Messner presso la University of New Mexico School of Medicine; essi hanno riferito di un aumento nell'incidenza di certi anticorpi anti-linfociti in individui conviventi con i pazienti affetti da LES, ma senza legami di parentela con loro. Anche altri studi compiuti da Schur, Kunkel, dall'autore e da Talal sugli anticorpi che reagiscono con l'RNA hanno sollevato la possibilità che siano coinvolti antigeni virali; tuttavia è difficile escludere il ruolo degli antigeni cellulari a favore di quelli virali.

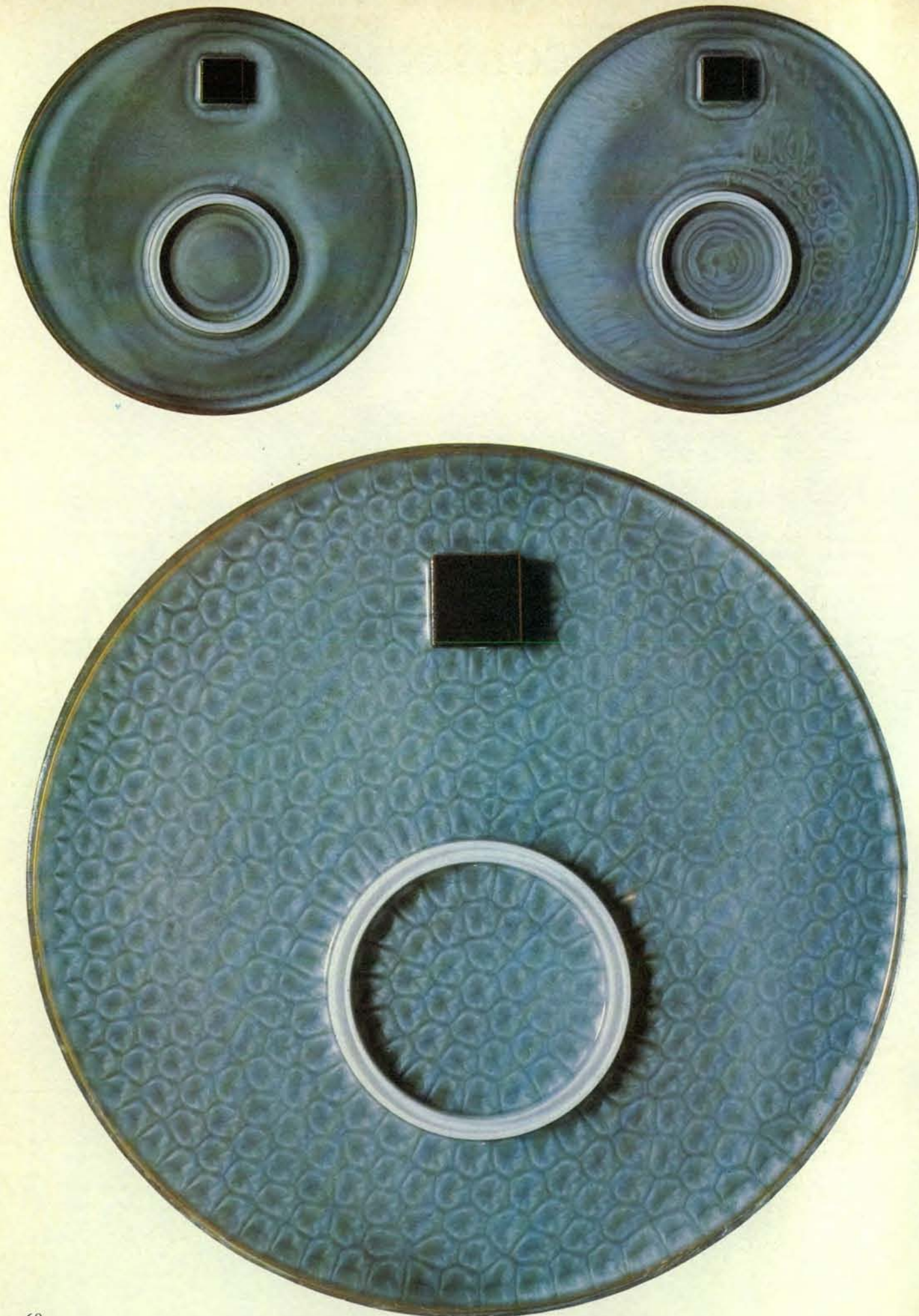
È chiaro che finora il meccanismo del danno ai tessuti è stato stabilito molto più chiaramente di qualsiasi fattore causale. Il massimo che si possa dire sull'eziologia è che la malattia probabilmente deriva dall'interazione di fattori genetici e uno o più altri fattori come farmaci, ormoni e virus. Per una comprensione più profonda delle cause del LES sarà necessario conoscere più approfonditamente le funzioni immunitarie normali.

Le migliorate possibilità di diagnosticare il LES in fase precoce, il chiarimento dei meccanismi del danno ai tessuti e un miglior controllo del decorso della malattia per mezzo delle analisi del sangue hanno reso più efficaci i metodi di cura, alleviando in questo modo i sintomi della malattia, particolarmente le manifestazioni renali. Più del 90 per cento dei pazienti sopravvivono per più di cinque anni dall'instaurarsi del LES, in confronto al 50 o 60 per cento di appena 25 anni fa. Il miglioramento della possibilità di diagnosi ha anche rivelato che questa malattia è molto più diffusa di quanto si pensasse.

Gli studi degli aspetti immunologici del LES hanno portato contributi fondamentali alla conoscenza di come gli agenti immunologici danneggino i tessuti. Nel LES evidentemente avviene qualcosa di irreparabile nell'ecosistema immunitario; la fragile simbiosi dei suoi vari elementi intercomunicanti - le cellule *B*, le cellule *T*, gli anticorpi, il sistema complemento e altri - viene distrutta. Le ricerche sul funzionamento disordinato del sistema immunitario nel LES dovrebbero contribuire a chiarire il «linguaggio» di comunicazione di questi elementi. Dovrebbero anche condurre a una comprensione più precisa delle aberrazioni immunologiche in altre malattie, come l'artrite reumatoide, la sclerosi a placche e la glomerulonefrite cronica.

La convezione

Lo spontaneo movimento verso l'alto di un fluido riscaldato può essere compreso soltanto analizzando le complesse relazioni tra temperatura, viscosità, tensione superficiale e ulteriori caratteristiche del fluido

di Manuel G. Velarde e Christiane Normand

La convezione dovrebbe essere familiare a chiunque abbia osservato l'intorbidamento di un brodo riscaldato, sentito il risucchio della cappa di un caminetto od osservato il luccicante tremolio delle correnti d'aria su una strada asfaltata in un giorno di sole. Lo stesso meccanismo convettivo è responsabile delle grandi correnti oceaniche e della circolazione globale dell'atmosfera; esso dà origine a moti su scala ancora maggiore nella fotosfera solare. Certi tipi di nubi si formano quando l'aria calda e umida sale verso l'alto in turbini convettivi ed è l'interruzione del normale trasporto per convezione che lascia periodicamente città come Los Angeles e Madrid sotto una cappa di smog durante un'inversione termica. Altri esempi sono meno familiari o meno facilmente osservati. La convezione ha un'influenza importante sull'asciugatura di pellicole di vernice e sulla dispersione di gas e di particelle nei polmoni. La convezione nel mantello terrestre è in apparenza la forza motrice della lenta migrazione dei continenti.

I più elementari tipi di convezione pare abbiano una semplice spiegazione che si può riassumere con l'espressione «salite di calore». Nei casi più semplici il flusso convettivo ha inizio quando un fluido (un gas o un liquido) viene riscaldato dal sotto. In seguito al riscaldamento lo strato inferiore del fluido si espande e diventa perciò meno denso degli strati sovrastanti. Lo strato inferiore più caldo e più leggero tende quindi a salire, mentre lo strato superiore più freddo tende a scendere in basso. Questo fenomeno era noto fin dal secolo XVIII. Può perciò apparire sorprendente che la formulazione di un dettagliato e quantitativo resoconto della convezione si sia rivelato una duratura sfida per l'ingegno dei teorici. In realtà, perfino di sistemi più semplici sottoposti a un intenso moto convettivo non si può dare una esatta descrizione matematica.

La natura delle difficoltà teoriche può essere esemplificata se si considera ancora il caso di uno strato fluido riscaldato dal sotto. La forza che guida il flusso convettivo in tale fluido è la spinta verso l'alto dello strato riscaldato e il modulo di tale forza è determinato dalla differenza di temperatura tra la superficie superiore e quella inferiore dello strato. La complessità dell'argomento appare chiara quando si scopre che la distribuzione di temperatura è notevolmente alterata dallo stesso flusso convettivo, che trasporta calore dalla superficie inferiore a quella superiore dello strato. In tal modo la forza che provoca il flusso è soggetta a modificazioni da parte dello stesso flusso.

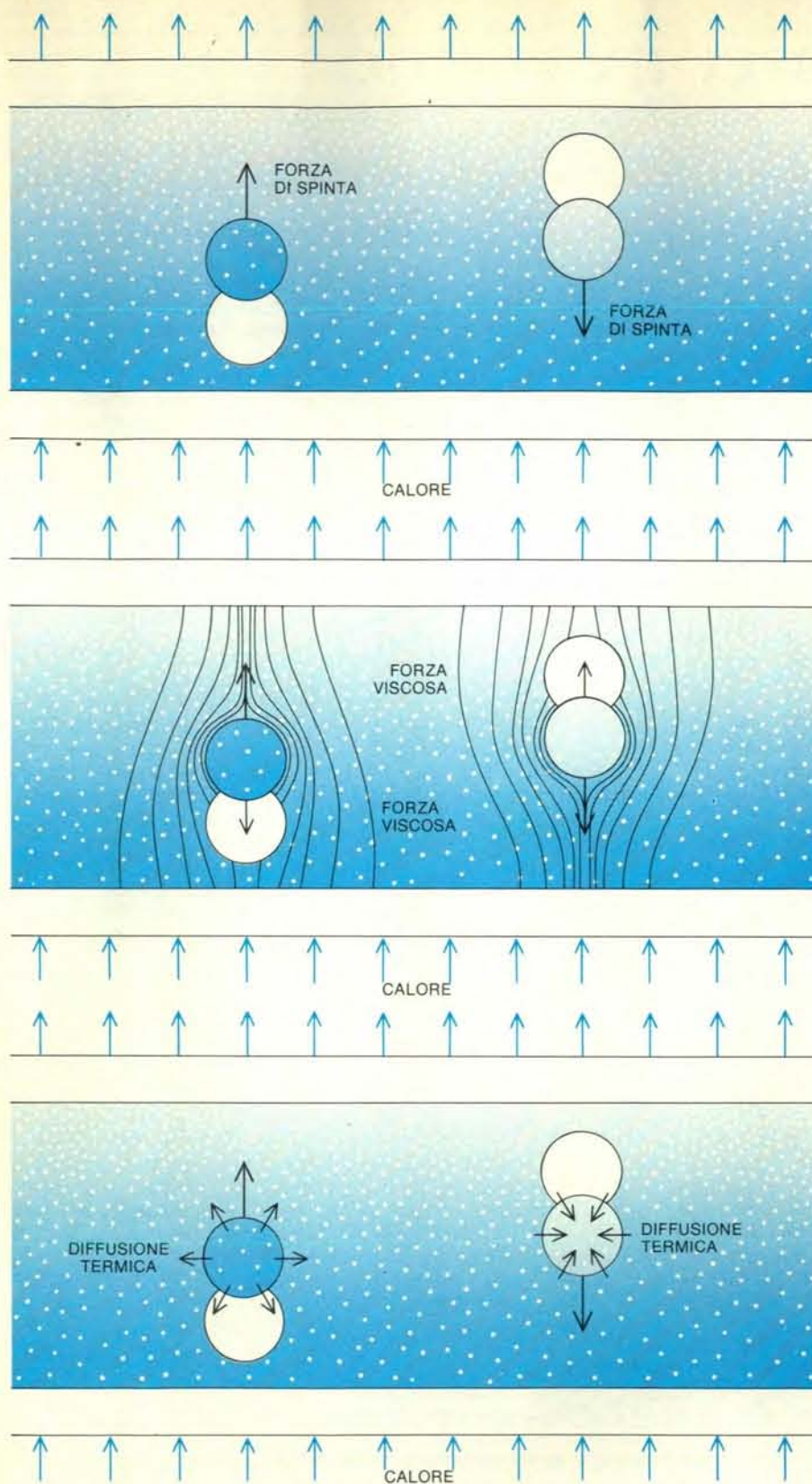
Pur se mancano ancora soluzioni esatte di problemi come questo, negli ultimi vent'anni sono stati compiuti sostanziali progressi verso una teoria generale della convezione. I progressi si sono avuti principalmente dall'adattamento di concetti e di tecniche matematiche sviluppati in altri campi delle scienze fisiche, più che altro nello studio delle transizioni di fase, dei materiali ferromagnetici e dei superconduttori. Questi metodi conducono a una analisi della stabilità delle varie modalità di moto nel fluido e quindi a previsioni della modalità più facilmente osservabile. I risultati sono soltanto approssimati, ma in alcuni casi l'approssimazione è buona, al punto da far sperare che sia possibile spiegare quello che succede in un brodo in ebollizione incipiente.

Il tipo di trasporto convettivo che tratteremo qui è chiamato convezione naturale o libera, che significa che il flusso è dovuto a forze che agiscono in seno al fluido. La forza è molto spesso la gravità, ma vi sono dei casi in cui giocano un ruolo significativo e persino fondamentale altri agenti, quali la tensione superficiale o un campo elettromagnetico. La convezione naturale è definita in contrapposizione alla convezione forzata, nella quale il moto del fluido è indotto da una forza impressa dall'esterno, per esempio da una pompa o da un ventilatore.

Una delle prime descrizioni della convezione naturale risale all'incirca al 1790 ed è dovuta a Benjamin Thompson, il conte Rumford. Egli avanzò l'idea che si poteva spiegare il trasporto di calore in una torta di mele. C'erano state precedenti proposte di un meccanismo convettivo per la circolazione atmosferica e durante il secolo XIX sono stati riportati in letteratura un certo numero di lavori aneddotici. Ricerche sistematiche sono iniziate tuttavia soltanto verso il 1900. Il lavoro sperimentale più significativo di tale periodo si deve al ricercatore francese Henri Bénard. Egli studiò un sistema convettivo più complesso delle sue stesse conoscenze e la sua vera natura è stata riconosciuta solo di recente. Le osservazioni di Bénard e la loro moderna interpretazione verranno riportate più avanti.

Il principale teorico della convezione dell'inizio del secolo XX è stato John William Strutt, alias Lord Rayleigh. Tra i suoi ultimi lavori c'era un articolo sulla convezione, pubblicato nel 1916, che tentava di spiegare i risultati di Bénard. Oggi sappiamo che la teoria di Lord Rayleigh non vale per il sistema esaminato da Bénard, tuttavia il lavoro di Lord Rayleigh è il punto di partenza per quasi tutte le moderne teorie della convezione.

La teoria di Lord Rayleigh può essere spiegata nel contesto di un modello sperimentale che impiega un fluido con pro-

Celle di convezione con una geometria poligonale caratteristica si formano spontaneamente quando viene riscaldato dal sotto un sottile strato di fluido. Nelle fotografie in alto nella pagina a fronte è riprodotta l'evoluzione del modello nel corso di alcune ore; la foto in basso mostra il modello completamente sviluppato. Inizialmente le celle sono lunghi «rulli» che seguono i contorni dello strato; i rulli danno origine a poligoni, che tendono a una forma di esagoni regolari e sono indipendenti dal contorno. In ciascuna cella il fluido sale nel centro e discende alla periferia. La circolazione è indotta principalmente da forze associate alla tensione superficiale e in molti fluidi assume questa forma solo quando la superficie superiore è libera. Le fotografie sono state riprese nel laboratorio di uno degli autori (Velarde) all'Università Autonoma di Madrid. Il fluido è un olio di silicone in cui sono stati sospesi fiocchi di alluminio per rendere visibile il flusso.

Per dare inizio a un flusso convettivo è necessario uno squilibrio di forze. Le forze si possono analizzare in un ipotetico esperimento nel quale un sottile strato di fluido tra due lastre rigide viene riscaldato dal sotto generando un gradiente di temperatura e di densità. Se viene spostata leggermente verso l'alto una porzione di fluido caldo nei pressi della parte inferiore dello strato, essa entra in una regione di maggiore densità media ed è perciò sottoposta a una forza di spinta verso l'alto. Allo stesso modo, se viene spostata verso il basso una porzione di fluido freddo nei pressi della parte superiore dello strato, essa diventa più pesante dell'ambiente circostante e tende a scendere verso il basso. La forza di spinta è ostacolata dalla resistenza viscosa e dalla diffusione del calore, che tende a uguagliare la temperatura di una porzione spostata con quella del suo ambiente circostante. L'importanza relativa di questi effetti è misurata dal numero di Rayleigh. La convezione ha inizio quando la spinta verso l'alto supera gli effetti dissipativi della resistenza viscosa e della diffusione termica, cioè quando il numero di Rayleigh supera un valore critico.

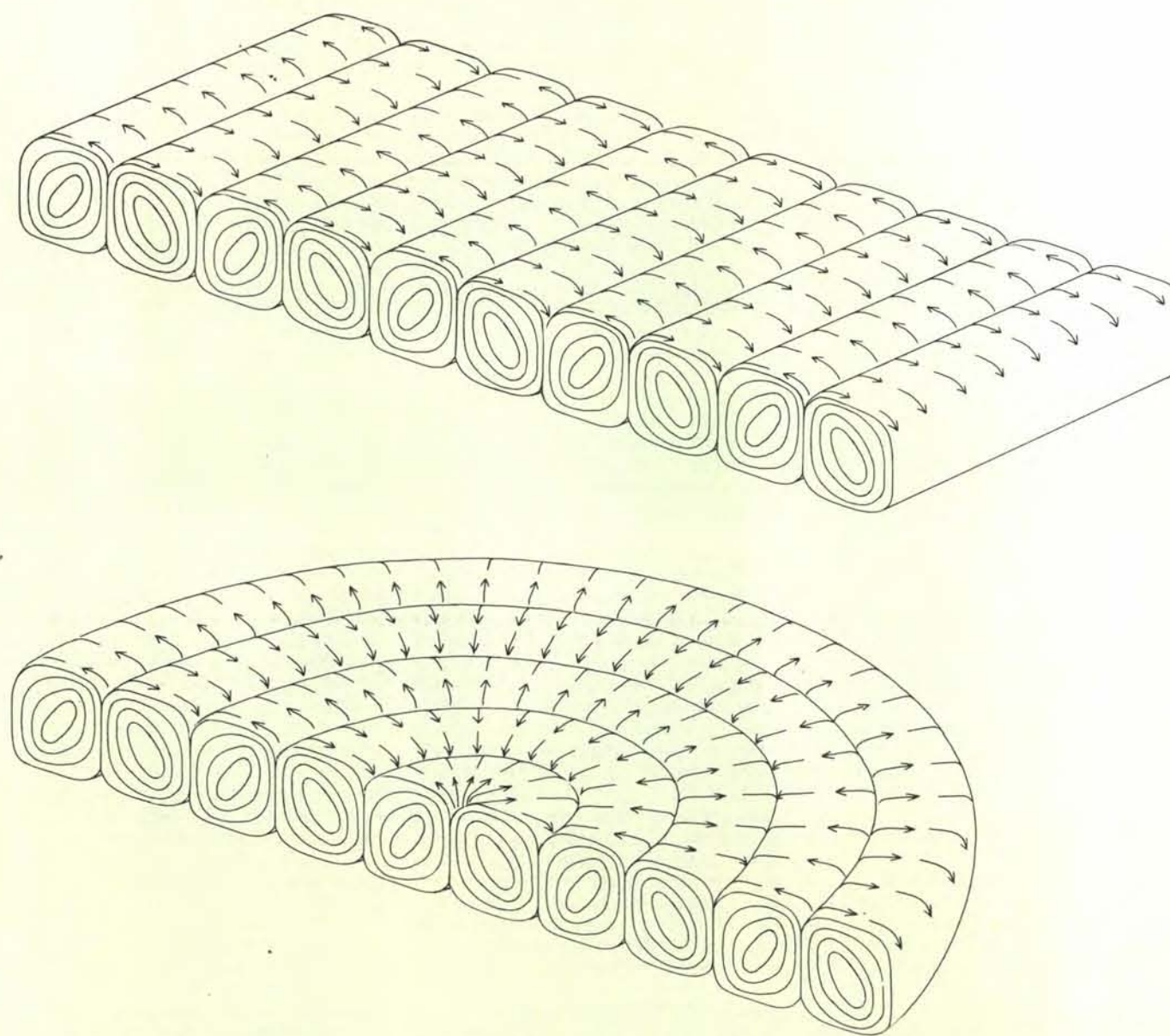
prietà alquanto più semplici di un qualsiasi gas o liquido reale. Un sottile strato di fluido viene confinato tra due lastre piane, rigide, orizzontali e riempie completamente lo spazio tra esse, in modo che non esista alcuna superficie libera. Per strato sottile si intende che l'estensione orizzontale dello strato è molto maggiore della sua profondità (che è uguale alla distanza tra le lastre). L'ipotesi di uno strato sottile serve per rendere minimi gli effetti al contorno delle lastre, che non entrano esplicitamente nella descrizione teorica. Lo strato ideale dovrebbe avere estensione orizzontale infinita; in pratica uno strato di qualche centimetro di larghezza e di pochi millimetri di spessore soddisfa ai requisiti.

L'apparecchiatura deve essere riscaldata dal sotto in modo che il fondo dello strato abbia una temperatura uniforme e costante. Allo stesso modo, viene fornito calore alla parte superiore dello strato in modo che anche là la temperatura sia costante e uniforme, ma più bassa di quella della superficie inferiore. Ne deriva, ovviamente, che anche la differenza di temperatura tra la superficie superiore e quella inferiore è costante e uniforme. Inoltre, si richiede che il gradiente di temperatura - la variazione di temperatura corrispondente a un dato spostamento verticale - sia lineare, cioè che un grafico della temperatura in funzione dell'altezza dia una linea retta.

A questo elenco vanno aggiunte poche altre ipotesi semplificative. Una è che la gravità sia la sola forza che agisce all'interno del fluido. Dato che gli esperimenti pratici sono su piccola scala, il campo gravitazionale dovrà essere praticamente uniforme in tutto il volume. Il fluido deve inoltre essere incomprimibile, ipotesi questa realistica per liquidi poco profondi. Infine, un aspetto molto importante del modello è che la variazione di temperatura alteri direttamente soltanto una proprietà del fluido. Questa proprietà è la densità, che diminuisce con l'aumentare della temperatura. In altri termini, il fluido riscaldato si espande, proprietà questa caratteristica dei gas e dei liquidi reali.

Esiste un metodo teorico ben definito per indagare sugli effetti del gradiente termico in questo esperimento ipotetico. Per prima cosa dobbiamo supporre che una porzione di fluido sia stata spostata verso l'alto o verso il basso rispetto alla sua posizione originale; quindi si devono analizzare le forze agenti sulla porzione spostata. Sono tali forze che determineranno tutti i successivi movimenti del fluido. La porzione può avere dimensioni e forma qualsiasi, ma lo spostamento deve essere piccolo. (La teoria di Rayleigh vale essenzialmente per spostamenti infinitesimali o di entità al limite infinitesime.) Lo spostamento iniziale non deve essere la conseguenza di una forza impressa; dal momento che le molecole di un fluido sono costantemente in movimento le loro posizioni fluttuano casualmente e ci si può aspettare che, se si attende abbastanza a lungo, qualsiasi spostamento piccolo avvenga spontaneamente. Si consideri una piccola porzione di fluido nei pressi della superficie inferiore dello strato. Per l'elevata temperatura sul fondo, la porzione ha una densità minore della densità media dell'intero strato. Però, mentre la porzione resta sul luogo viene circondata da fluido della stessa densità e quindi la sua spinta verso l'alto è nulla. Tutte le forze che agiscono su di essa sono equilibrate ed essa non sale verso l'alto e non affonda.

Supponiamo ora che per effetto di qualche perturbazione casuale alla porzione di fluido venga impresso un leggero moto verso l'alto. Quale effetto ha lo spostamento sull'equilibrio delle forze? La porzione è ora circondata da un fluido più freddo e più denso. Il risultato è che su essa agisce una spinta verso l'alto positiva che la fa salire. La forza complessiva verso l'alto è proporzionale alla differenza di densità e al volume della porzione. Così uno spostamento iniziale verso l'alto del fluido caldo viene amplificato dal gradiente di densità e questo effetto dà origine a forze che provocano un ulteriore movimento verso l'alto. Un'analisi del genere si potrebbe fare per un lieve spostamento verso il basso di una porzione di fluido freddo e denso nei pressi della parte alta dello strato. Muovendosi verso il basso la porzione entrerebbe in un ambiente di densità media più bassa, diventando più pesante dell'ambiente circostante. Essa tende perciò ad affondare, amplificando la perturbazione iniziale. La convezione naturale è il risultato di questi flussi combinati verso l'alto e verso il basso e tende a rovesciare l'intero strato di fluido. La conseguenza di questa analisi è che si dovrebbe osservare la convezione nel fluido tutte le volte che esiste un gradiente termico, per quanto piccolo esso sia. Anche in un gradiente infinitesimo, qualsiasi moto casuale verso l'alto del fluido caldo o verso il basso del fluido freddo dovrebbe essere sufficiente per dare origine a un flusso. In realtà non si osserva tale estrema sensibilità, anzi il gradiente termico deve raggiungere un certo valore di soglia perché inizi il flusso convettivo. La spiegazione del perché le cose stanno così è stato il più significativo contributo di Lord Rayleigh.

Lord Rayleigh mise in evidenza che una teoria della convezione deve tener conto di almeno altri due effetti sul moto di una particella fluida. Uno di essi è la resistenza viscosa, l'equivalente dell'attrito nei fluidi. La resistenza viscosa è sempre opposta al moto del fluido e il suo modulo è determinato in parte da una proprietà intrinseca del fluido, la viscosità di taglio, che misura la resistenza al moto relativo di due qualsiasi regioni adiacenti.

Le celle a forma di rullo costituiscono una configurazione stabile nella convezione indotta da forze di spinta anziché dalla tensione superficiale. L'unità fondamentale del modello comprende due rulli ruotanti in verso opposto; la larghezza di tale unità è doppia della profondità dello strato fluido. La forma del modello dipende dai contorni dello strato. In un recipiente rettangolare i rulli sono paralleli ai lati più corti, mentre in uno cilindrico formano anelli concentrici. Si osserva un modello di rulli stabile soltanto quando il fluido non presenta superficie libera.

Se la velocità non è troppo grande, il modulo della resistenza viscosa è proporzionale alla viscosità moltiplicata per il raggio della porzione e la velocità. Ovviamente, se la resistenza viscosa è uguale alla spinta verso l'alto, non c'è moto.

Il secondo effetto dissipativo costituisce una conseguenza del fatto che la convezione non è il solo meccanismo di trasporto di calore in un fluido. Anche l'irraggiamento e la conduzione, o diffusione del calore, possono trasportare calore. Alle temperature relativamente basse della maggior parte degli esperimenti di convezione, l'irraggiamento fornisce un contributo talmente piccolo che può essere ignorato. Invece la diffusione del calore non è sempre trascurabile; essa tende ad annullare il gradiente termico che origina il flusso convettivo.

L'effetto della diffusione del calore può essere spiegato considerando ancora una porzione di fluido caldo spostata verso l'alto dalla sua posizione di equilibrio in un ambiente più freddo. Secondo la definizione fondamentale di calore, le molecole della porzione calda devono avere una velocità media maggiore di quelle del fluido circostante più freddo. Le molecole possono attraversare liberamente il contorno che definisce la porzione e l'effetto di molti scambi del genere in entrambe le direzioni sarà di uguagliare le velocità medie delle due popolazioni. In altre parole, il calore fuoriuscirà dalla porzione calda spostata in modo da farla raffreddare, mentre il fluido circostante si riscalda, fin quando alla fine essi raggiungono l'equilibrio alla stessa temperatura. Per una porzione di fluido freddo spostata verso il basso, il flusso di calore avviene in verso opposto, dall'ambiente caldo alla porzione fredda. In entrambi i casi la differenza di temperatura locale si riduce e allo stesso modo si comporta la spinta che da essa ha origine.

Il tempo necessario perché una porzione fluida raggiunga l'equilibrio termico con il suo ambiente circostante dipende in parte da una proprietà del fluido, la diffusività termica. La scala dei tempi di questo processo è inversamente proporzionale alla costante di diffusività ed è direttamente proporzionale alla superficie della porzione. Se questo tempo di diffusione termica è confrontabile con il tempo necessario perché la porzione si sposti di una certa distanza caratteristica, quale può essere il suo diametro, la spinta verso l'alto viene annullata. In altre parole, se il fluido si muove non più velocemente di quanto perde calore per diffusione, non può mantenersi un flusso convettivo. La fornitura di calore al sistema attraverso la lastra inferiore viene in tal caso effettuata attraverso lo strato fluido mediante un meccanismo puramente conduttivo o diffusivo, senza alcun moto di assieme.

L'analisi di Lord Rayleigh mostra che la semplice esistenza di un gradiente termico non è sufficiente a garantire l'inizio di un flusso convettivo. È necessario che la spinta verso l'alto risultante da tale gradiente vinca gli effetti dissipativi della resistenza viscosa e della diffusione del calore. L'energia potenziale gravitazionale liberata dalla discesa del fluido più denso e dalla risalita di quello più leggero deve essere maggiore dell'energia dissipata per viscosità e diffusione. La relazione tra tali effetti può essere espressa come un rapporto adimensionale: la forza di spinta divisa per il prodotto tra la resistenza viscosa e la velocità di diffusione del calore. Il rapporto è adimensionale nel senso che tutte le unità di misura associate alle tre grandezze si eliminano esattamente, lasciando un numero puro il cui valore è lo stesso qualunque sia il sistema di unità adottato. Tale rapporto è oggi chiamato numero di Rayleigh. La convezione ha inizio quando il numero di Rayleigh supera un certo valore critico.

Il significato del numero di Rayleigh può essere riaffermato con maggiore esattezza attraverso un esame della stabilità di varie possibili modalità di moto nel fluido. È conveniente definire la stabilità in funzione di una curva di potenziale o di una superficie di potenziale, che esprime l'energia di un sistema in funzione di qualche variabile. Il sistema si trova solitamente in un qualsiasi stato che abbia minima energia, corrispondente al punto più basso sulla superficie di potenziale.

È abbastanza facile immaginare un modello reale di una superficie di potenziale, cioè una boccia emisferica contenente una biglia. In condizioni di equilibrio la biglia è in quiete sul fondo della boccia, dove la sua energia potenziale gravitazionale è minima. Se qualche perturbazione casuale sposta adesso leggermente la biglia, essa rotolerà indietro verso la sua posizione di equilibrio; può darsi che essa superi il punto più basso della superficie e oscilli attorno a esso, ma gli effetti dissipativi dell'attrito finiranno con lo smorzare le oscillazioni e dopo un certo tempo la biglia si ritroverà in quiete nel punto di minima energia. Dato che una biglia sul fondo della boccia ritorna alla sua posizione originale dopo una perturbazione essa è in equilibrio stabile.

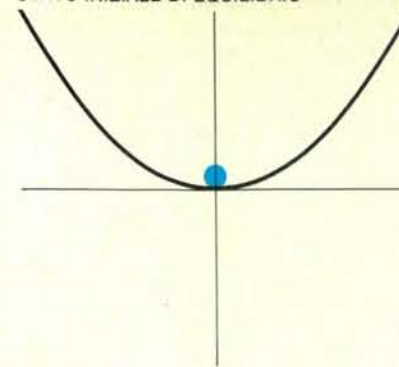
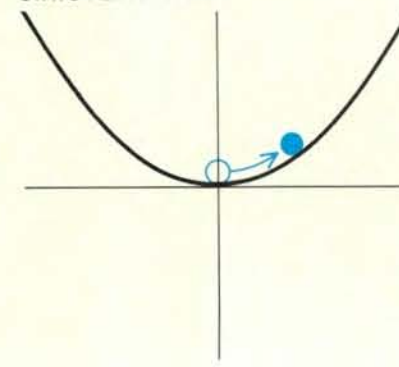
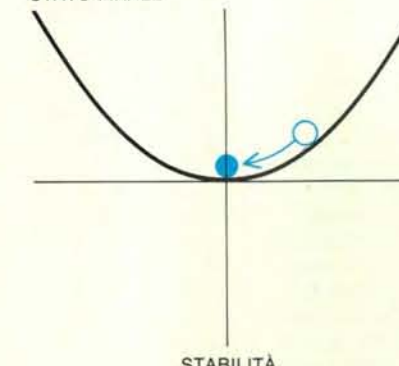
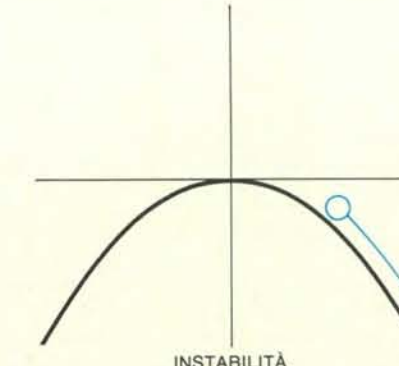
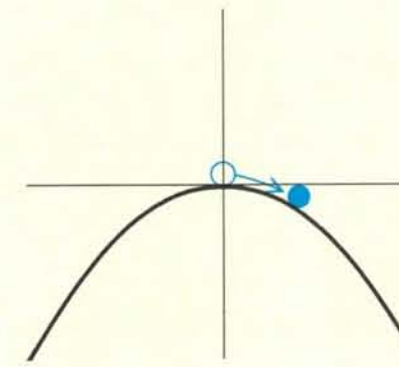
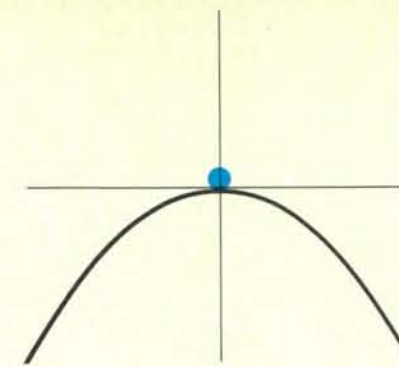
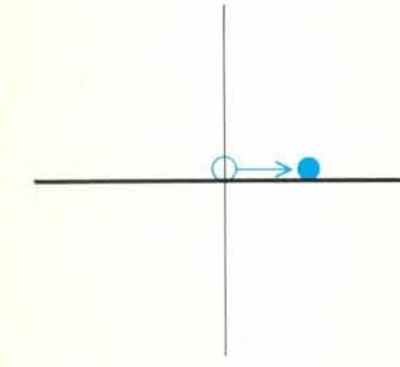
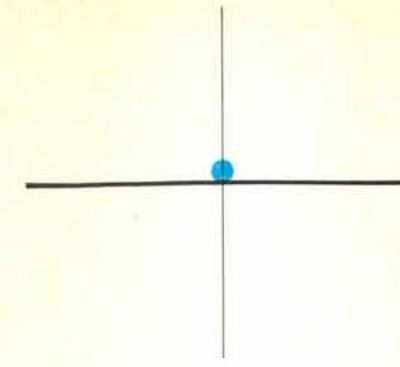
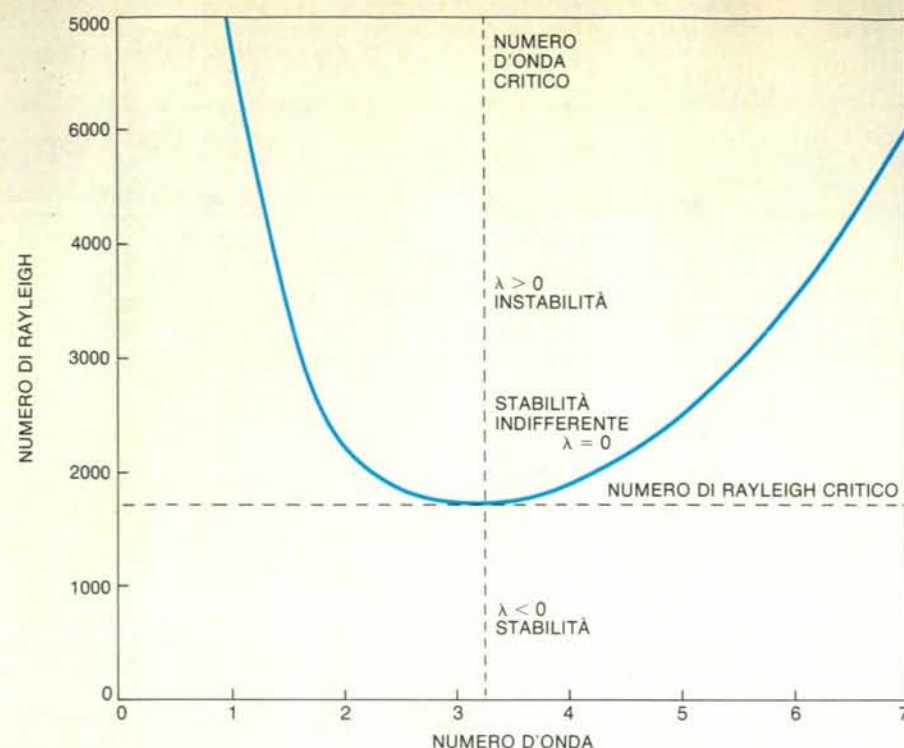
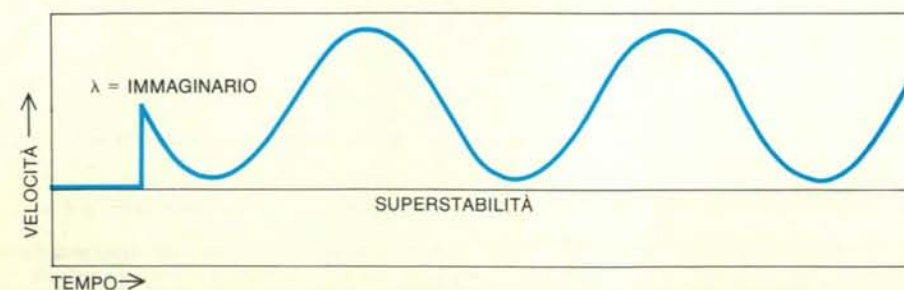
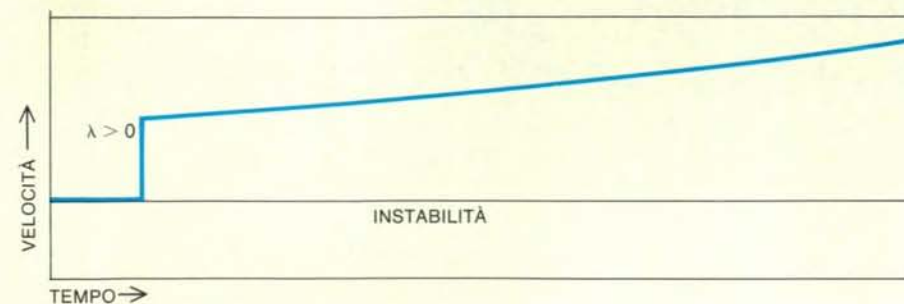
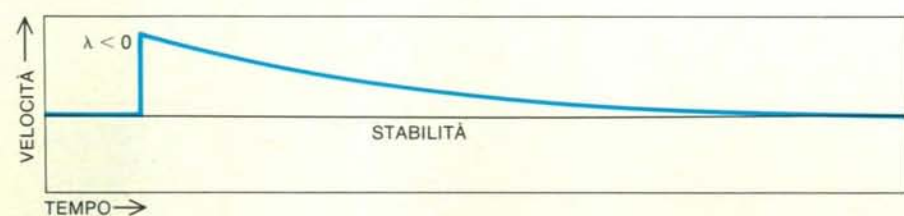
Un altro modello di superficie di potenziale si ricava capovolgendo la boccia emisferica e disponendo accuratamente in equilibrio la biglia sul suo vertice. Anche questo è uno stato di equilibrio, nel senso che tutte le forze agenti sulla biglia sono equilibrate, e in assenza di qualsiasi perturbazione, la biglia potrebbe restare indefinitamente in quiete. In pratica, però, ci sarà sempre qualche influenza esterna (per esempio, una corrente d'aria o un camion in transito) che disturberà questo equilibrio precario se lo sperimentatore aspetta abbastanza a lungo. Dopo una siffatta perturbazione la biglia non ritorna al punto di equilibrio ma se ne allontanerà indefinitamente. Per quanto piccola sia la perturbazione iniziale, la biglia si troverà alla fine a una grande distanza dal centro. La perturbazione è amplificata e quindi lo stato di equilibrio è detto instabile.

C'è una terza possibilità: la biglia può anche essere posta su una superficie piana. In questo caso, se spostiamo la biglia essa non ritorna nella sua posizione originale né continua ad allontanarsi da essa, ma si limita a restare in equilibrio nella sua nuova posizione. Si dice che ogni punto di una superficie di potenziale piana rappresenta uno stato di equilibrio indifferente o marginale.

Se meditiamo un po' di più su questo modello di superficie di potenziale si conclude che la stabilità assoluta di un sistema può essere dimostrata soltanto verificando la sua risposta a tutte le perturbazioni possibili. Per esempio, una biglia all'interno della boccia ritornerà al centro dopo una perturbazione infinitesima o anche dopo una piccola, ma finita. Non ritornerà invece al suo punto di partenza se la perturbazione è sufficientemente grande da scagliarla completamente all'esterno della boccia. Poiché si dovrebbero verificare un numero infinito di possibili perturbazioni, è difficile dimostrare se uno stato di equilibrio è stabile; invece l'instabilità si può dimostrare mettendo in evidenza anche una sola perturbazione che cresce spontaneamente.

L'applicazione di questi principi al problema della convezione è immediata. Si dimostra facilmente che uno strato di fluido immobile riscaldato uniformemente dal sotto è in uno stato di equilibrio, anche se si trascurano lo smorzamento viscoso e la diffusione del calore. Anche se la porzione più leggera del fluido è sovrastata da materiale più denso, in modo da poter diminuire l'energia potenziale gravitazionale scambiando le due posizioni, in assenza di perturbazioni tutte le forze che agiscono su una qualsiasi porzione di fluido sono in equilibrio. La domanda alla quale deve rispondere una teoria della convezione è se l'equilibrio è stabile o instabile o se presenta invece una stabilità indifferente. In altre parole, la teoria deve definire la forma della superficie di potenziale.

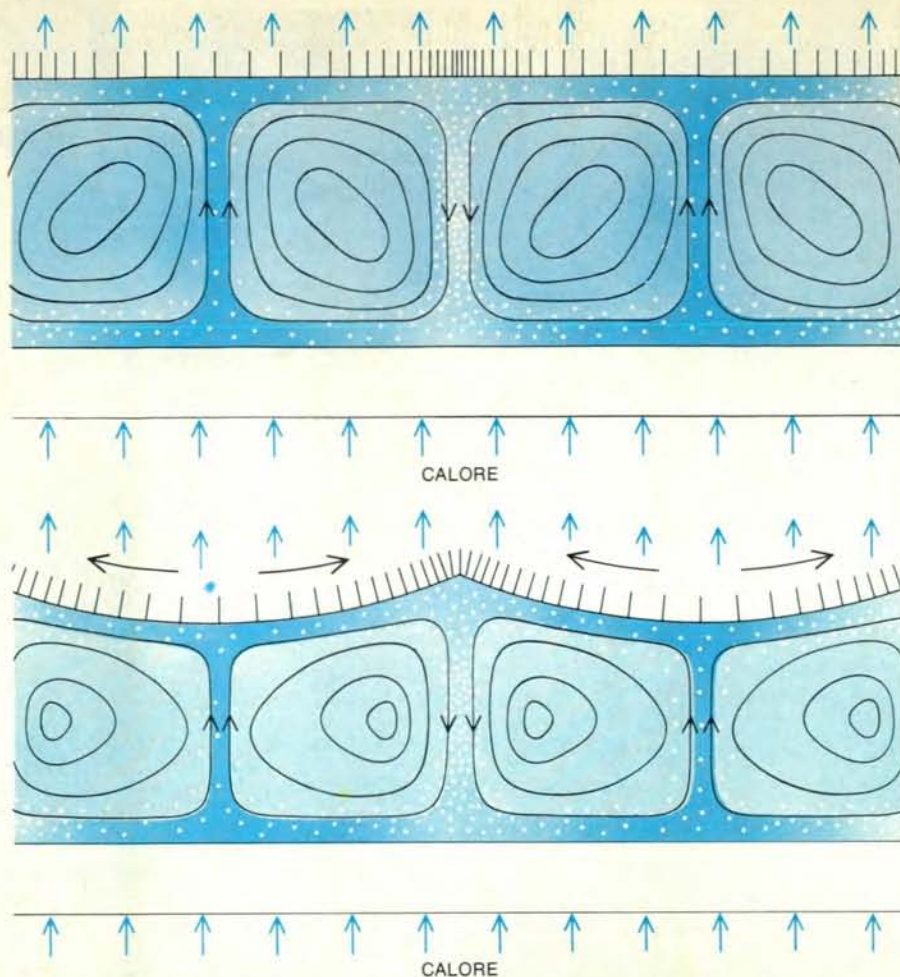
È il valore del numero di Rayleigh a determinare la curvatura della superficie di potenziale. Se il numero di Rayleigh è nullo, per esempio perché sono nulli il gradiente termico e la forza di spinta, lo stato di quiete è ovviamente stabile e la superficie di potenziale ha la concavità verso l'alto, come l'interno di una boccia o di una valle. Per mettere in moto il fluido è necessario aumentarne l'energia. D'altra parte, se il numero di Rayleigh è molto grande, in modo tale che la spinta verso l'alto superi tutti gli effetti dissipativi, il fluido può diminuire la sua energia totale dando luogo a un flusso convettivo. In tal caso qualsiasi perturbazione dell'equilibrio stazionario verrà amplificata. La superficie di potenziale ha la concavità verso il basso come una boccia capovolta.

La continuità impone che esista qualche valore del numero di Rayleigh, intermedio tra questi estremi, in corrispondenza del quale la spinta verso l'alto e le forze dissipative abbiano lo stesso modulo. È questo il numero di Rayleigh critico, che caratterizza la superficie di potenziale piana, di transizione o di stabilità indifferente. Quando il numero di Rayleigh

La stabilità di un sistema fisico si può giudicare dalla sua risposta a una perturbazione arbitraria, quale un piccolo spostamento di una biglia in quiete su una superficie. Se la superficie è concava, la biglia ritorna alla posizione di equilibrio sul fondo della boccia: tale posizione è detta di equilibrio stabile. Su una superficie convessa la biglia può essere tenuta in equilibrio sul vertice, ma qui l'equilibrio è instabile: la più piccola perturbazione è amplificata come la biglia diminuisce la propria energia potenziale rotolando verso il basso. Su una superficie piana la biglia non ritorna nella sua posizione di partenza né si allontana ulteriormente da essa: si parla di equilibrio indifferente. Un sistema può essere stabile di fronte a certe perturbazioni, ma non ad altre, come per una superficie che presenta sia regioni concave sia convesse. In un fluido semplice per innescare la convezione è necessaria una distribuzione instabile di alcune proprietà quali la densità o la tensione superficiale.

aumenta a partire da zero (per esempio, a causa di un crescente gradiente termico) la superficie di potenziale parte concava e gradualmente si appiattisce; in corrispondenza del numero di Rayleigh critico, essa è perfettamente piana, mentre, quando il numero continua ad aumentare, la superficie diventa convessa. L'equilibrio di uno stato di quiete diventa instabile soltanto quando si supera il valore critico. Per l'esperimento modello qui considerato, i calcoli indicano un numero di Rayleigh critico di 1708. In un esperimento di laboratorio, dove il fluido è uno strato di olio di silicone di qualche millimetro di spessore, il numero di Rayleigh critico viene raggiunto quando il gradiente termico è di pochi gradi centigradi.

L'esperimento ipotetico su cui si basa questa spiegazione della teoria di Lord Rayleigh comprende molte ipotesi semplificative, alcune delle quali contrarie ai fatti. La teoria ha comunque avuto un notevole successo nel predire le condizioni necessarie per l'inizio della convezione nei fluidi reali. Per esempio, gli esperimenti di Peter L. Silveston dell'Università della Columbia Britannica e di Ernest L. Koschmieder dell'Università del Texas ad Austin hanno fornito valori del numero di Rayleigh critico di 1700 ± 50, in buon accordo con il valore teorico.

L'equilibrio tra la spinta verso l'alto e le forze dissipative fornisce un criterio per l'innesco dell'instabilità convettiva, ma che cosa si osserva nell'esperimento modello una volta iniziato il flusso? La teoria di Rayleigh fornisce soltanto indicazioni limitate nel rispondere a tale domanda e anche le teorie più complesse che discuteremo in seguito non sono in grado di tener conto di tutti gli aspetti osservati in una circolazione convettiva pienamente avviata. Tuttavia, anche se non si può ricavare matematicamente l'evoluzione del flusso, è almeno possibile una descrizione qualitativa.

In uno strato fluido riscaldato uniformemente dal sotto, il gradiente termico dovrebbe essere indipendente dalla posizione orizzontale e la stessa cosa dovrebbe valere per la spinta verso l'alto risultante. Quando si supera il numero di Ray-

La stabilità di un fluido di fronte alla convezione indotta dalla spinta verso l'alto viene determinata dal valore del numero di Rayleigh. Se il numero è inferiore al valore critico, qualsiasi perturbazione viene smorzata, mentre se il numero supera il valore critico, la perturbazione continua a crescere. Il sistema è più sensibile a perturbazioni con un particolare numero d'onda, o scala di lunghezza, corrispondente al doppio della profondità dello strato fluido. La differenza tra il numero di Rayleigh effettivo e il numero critico determina un parametro λ; nella teoria proposta da Rayleigh la velocità del flusso convettivo dipende da λ in modo esponenziale. Se λ è negativo, la velocità scende a zero; se λ è nullo la velocità resta costante; se λ è positivo, la velocità continua ad aumentare. Dato che λ è un numero complesso che può avere un valore immaginario, il flusso può oscillare, fenomeno detto superstabilità.

Le variazioni di tensione superficiale modificano il modello del trasporto convettivo in un fluido dotato di superficie libera. Il valore della tensione superficiale varia al variare della temperatura, ed è maggiore quando il fluido è più freddo. Qualsiasi differenza di temperatura attraverso la superficie dà quindi origine a un gradiente di tensione superficiale. Lo stato di quiete diventa instabile se il gradiente è abbastanza grande da vincere gli effetti dissipativi della viscosità e della diffusione termica. Il valore della tensione viene in questo caso rappresentato dalla intensità del colore. Il fluido viene trascinato lungo la superficie verso regioni più fredde caratterizzate da maggiore tensione superficiale e viene sostituito da fluido più caldo proveniente dal basso.

leigh critico e l'equilibrio in stato di quiete diventa instabile, il fluido caldo tende a salire verso l'alto dovunque, mentre quello freddo tende a discendere dovunque. Ovviamente i due processi non possono verificarsi contemporaneamente. Questo ostacolo viene evitato dalla divisione spontanea dello strato in un modello di «celle» di convezione, in ognuna delle quali il fluido circola in un'orbita chiusa. Argomentazioni teoriche danno un'indicazione approssimativa della scala favorita dei singoli elementi nel modello di convezione. Queste argomentazioni derivano dalla sensibilità variabile dello stato marginalmente stabile a perturbazioni con differenti scale di lunghezza. Qui dobbiamo stare attenti a non confondere l'ampiezza di una perturbazione, che nell'esperimento modello corrisponde allo spostamento verticale di una porzione di fluido, con la scala della perturbazione, che misura le dimensioni della porzione. Se la teoria di Rayleigh deve fornire risultati significativi, l'ampiezza deve sempre essere infinitesima, mentre la scala può raggiungere le dimensioni massime consentite dall'apparecchiatura.

Si è soliti esprimere la scala di una perturbazione in termini di un numero d'onda, che è il reciproco di una lunghezza. Questa abitudine riflette il fatto che la geometria di una perturbazione è generalmente complessa, in modo tale da non avere una sola dimensione chiaramente definita; la perturbazione può essere però decomposta in uno spettro di modi fondamentali, o frequenze spaziali, proprio come un suono complesso può essere analizzato in una combinazione di note pure. Un numero d'onda rappresenta i contributi di una particolare scala di lunghezza alle fluttuazioni casuali. Numeri d'onda maggiori corrispondono a scale più piccole.

La stabilità dello stato stazionario in quiete è suscettibile di essere sconvolta da perturbazioni con certi numeri d'onda piuttosto che con altri. Si può immaginare un esperimento in cui venga misurato il numero di Rayleigh critico in un fluido le cui fluttuazioni possono essere controllate in modo che siano sempre caratterizzate da un solo numero d'onda. Un siffatto esperimento dovrebbe rivelare che l'instabilità si instaura più rapidamente quando il numero d'onda descrive perturbazioni con una dimensione orizzontale circa doppia della profondità del fluido. Per numeri d'onda maggiori o minori, sono necessarie condizioni più estreme (un maggior numero di Rayleigh) per indurre la convezione. Il valore calcolato 1708 per il numero di Rayleigh critico è quello trovato quando le fluttuazioni hanno le dimensioni più adeguate.

La sensibilità del fluido a perturbazioni su una particolare scala implica che tali perturbazioni saranno amplificate più rapidamente di quasiasi altra quando lo strato diventa instabile. Pertanto il modello osservato quando ha inizio la convezione potrebbe presentare qualche peculiarità in prossimità di tale scala. Non è certo che tali peculiarità persistano una volta pienamente sviluppato il flusso convettivo, ma risulta che ciò avviene, purché il numero di Rayleigh non superi di troppo il valore critico.

Il numero d'onda precisa la scala complessiva del modello ma non la sua forma dettagliata; si potrebbero costruire celle di convezione di molte forme differenti in modo che abbiano lo stesso numero d'onda. Il modello osservato in realtà dipende fortemente dalla geometria dell'apparecchiatura sperimentale; esso non può essere ricavato dai principi primi, ma esistono regole empiriche ben consolidate che fanno previsioni qualitative attendibili.

In esperimenti quale quello ipotetico ora descritto, nel quale le superfici superiore e inferiore sono vincolate da contorni rigidi, l'elemento fondamentale del modello è un «rullo» che ha una lunga forma tubolare. Il fluido caldo sale lungo un bordo del rullo, attraversa la superficie superiore perdendovi calore e ricade poi sul fondo dello strato lungo il bordo opposto. La circolazione trasporta quindi il fluido attraverso il contorno inferiore, dove la sua temperatura aumenta nuovamente. Rulli adiacenti hanno versi di rotazione opposti.

Quando si osserva un rullo in sezione trasversale, la sua forma è all'incirca quadrata: la larghezza è uguale all'altezza e quest'ultima è ovviamente determinata dalla profondità dello strato fluido. In tal modo le proporzioni del rullo sono costanti, ma le sue dimensioni dipendono dalla profondità del fluido. Dato che l'unità ricorrente in questo modello è formata da due rulli ruotanti in senso opposto, la scala del modello è adeguatamente definita dalla larghezza di due rulli. Quindi la scala è uguale al doppio della profondità del fluido, secondo le previsioni dell'analisi con numeri d'onda.

La pianta del modello a celle (il suo aspetto dall'alto) è determinata in larga misura dalla forma dell'apparecchiatura. Dato che nella teoria non entrano direttamente dettagli di questo genere non si può facilmente ricavare la pianta; tuttavia l'osservazione sperimentale fornisce un'ampia base di previsione. Se il recipiente è rettangolare, i rulli tendono ad

allinearsi parallelamente ai lati più corti. La larghezza di ogni rullo e quindi il loro numero sono determinati dalla profondità dello strato. In un recipiente cilindrico i rulli formano anelli concentrici.

Come abbiamo detto sopra, l'analisi della convezione di Lord Rayleigh è stata ispirata principalmente dalle osservazioni sperimentali fatte da Bénard verso il 1900. Oggi sappiamo che la teoria di Rayleigh non è adatta per il meccanismo convettivo studiato da Bénard. Le condizioni sperimentali impiegate da Bénard differivano di poco ma in modo sostanziale da quelle qui descritte e l'importanza delle variazioni appare subito nel modello del flusso convettivo. Nella convezione di Bénard i rulli imposti dalla geometria dell'apparecchiatura possono apparire per breve tempo all'inizio del flusso, ma danno subito luogo a una figura più complessa: un mosaico poligonale della superficie fluida. Inizialmente i poligoni sono alquanto irregolari, avendo da quattro a sette lati, con numero medio di sei. Quando la figura si sviluppa pienamente, diventa uno schieramento quasi perfetto di esagoni regolari, disposti come in un alveare. Il centro di ogni cella esagonale è una regione di fluido caldo in movimento verso l'alto, che si estende sulla superficie superiore e ridiscende sul perimetro, dove le celle adiacenti sono unite.

Negli esperimenti di Bénard, come nell'esperimento ipotetico descritto in precedenza, il fluido forma uno strato sottile riscaldato dal sotto. La differenza essenziale è che lo strato non è confinato tra due contorni rigidi, ma è invece esposto all'aria sulla sua superficie superiore. Essendo la superficie libera, il flusso può essere influenzato dalla tensione superficiale; in effetti, è oggi nota l'influenza predominante della tensione superficiale nella convezione di Bénard, dove è più importante della forza di spinta. Non deve pertanto sorprendere che la teoria di Rayleigh, nella quale si assume esplicitamente che la spinta verso l'alto sia la sola forza agente, sia inadeguata a spiegare una convezione di questo tipo. Le previsioni sono errate perfino per grandezze fondamentali quale il gradiente termico necessario per innescare il flusso convettivo. Una valida teoria alternativa venne introdotta solo nel 1958 da J. R. A. Pearson dell'Imperial College of Science and Technology di Londra.

La tensione superficiale è la forza di coesione il cui effetto complessivo è quello di rendere minima la superficie di un fluido. Per esempio, una goccia libera di liquido tende ad assumere una forma sferica per effetto della tensione superficiale, essendo quella la configurazione di minima area. La tensione si può pensare come un intreccio di elastici tesi in tutte le direzioni sulla superficie libera. Se in un punto qualsiasi le forze esercitate dai vari elastici non si equilibrano, lo strato superficiale migrerà verso la regione di maggiore tensione fino a raggiungere l'equilibrio. Il flusso superficiale viene trasmesso all'intera massa fluida come risultato della sua viscosità di taglio.

La tensione superficiale può agire come forza propulsiva in un flusso convettivo perché la tensione varia con la temperatura: come la densità, la tensione superficiale diminuisce al crescere della temperatura. Qualsiasi gradiente termico instaurato attraverso la superficie liquida sarà accompagnato da un gradiente di tensione superficiale. Le regioni più fredde presenteranno una maggiore tensione superficiale, mentre in quelle più calde la tensione diminuirà. Se il gradiente di tensione superficiale porta a uno squilibrio delle forze, ne risulterà un flusso.

La forma a mosaico della superficie con celle esagonali è un aspetto caratteristico della convezione indotta da un gradiente di tensione superficiale. Quando la tensione è maggiore, la superficie si raggrinza e l'area diminuisce. Sopra al vertice ascendente nel centro di ogni cella la superficie presenta una depressione; il fluido deve risalire prima di discendere dal bordo della cella.

L'innesco dell'instabilità convettiva nel sistema di Bénard può essere analizzato allo stesso modo in cui si analizza l'innesco di un flusso indotto dalla spinta verso l'alto. Supponiamo che una porzione di liquido caldo venga spostata verso l'alto da qualche fluttuazione casuale. Che tale moto sia mantenuto o meno da forze di spinta, esso avrà comunque un effetto sulla superficie dello strato, facendo aumentare lievemente la temperatura e riducendo perciò la tensione superficiale nell'area direttamente al di sopra della fluttuazione. Le forze sulla superficie rimangono comunque in equilibrio perché la superficie circostante esercita una trazione uguale in tutte le direzioni su tale regione. Per poter dare inizio al flusso è necessaria una seconda perturbazione in grado di provocare uno spostamento orizzontale di qualche piccolo elemento della superficie nella zona di minor tensione. Le forze di tensione che agiscono sull'elemento spostato sono allora non equilibrate e, se il gradiente di tensione superficiale è abbastanza grande, lo spostamento sarà amplificato. L'elemento di superficie sarà trascinato nella regione più fredda di maggiore tensione e trascinerà con sé parte della massa fluida. Dagli strati più caldi sottostanti verrà quindi risucchiato verso l'alto più liquido, aumentando i gradienti superficiali di temperatura e di tensione. Nel frattempo il fluido che si è raffreddato durante l'attraversamento della superficie comincerà a scendere e si formerà il modello cellulare.

Come nella convezione indotta dalla spinta verso l'alto, la semplice esistenza di un gradiente termico non garantisce il mantenimento di un flusso convettivo. Il gradiente deve essere abbastanza grande da vincere gli effetti dissipativi della resistenza viscosa e della diffusione del calore. L'equilibrio di tali effetti nel sistema di Bénard viene espresso da un altro rapporto adimensionale, chiamato in questo caso con il nome del ricercatore italiano del XIX secolo C. G. M. Marangoni. La formula è la stessa del numero di Rayleigh con la differenza che la forza di spinta è sostituita dalla forza di tensione superficiale; in altre parole il numero di Marangoni è il rapporto tra il gradiente di tensione superficiale e il prodotto della resistenza viscosa per la diffusione del calore. La convezione di Bénard ha luogo quando il numero di Marangoni supera un valore critico.

Un aspetto caratteristico della convezione guidata da gradienti di tensione superficiale è che essa altera il contorno della superficie. Le regioni di elevata tensione superficiale tendono a raggrinzarsi diminuendo in tal modo la loro superficie totale esposta. Le conseguenze di questo effetto vanno contro l'intuito. Nel centro di una cella di Bénard, dove il fluido sta salendo, la superficie presenta una depressione, mentre ai bordi della cella, dove il fluido sta scendendo, la superficie è rialzata. Le forze gravitazionali si oppongono alla formazione di sporgenze della superficie, dal momento che l'energia potenziale gravitazionale è minima su

Un vertice ascendente di fluido caldo crea un gradiente di tensione superficiale lungo la superficie di separazione di due liquidi. La depressione della superficie indica il punto di minima tensione e il fluido viene trascinato lungo la superficie di separazione verso destra e lasciato lontano da tale punto. La circolazione è indotta dalla trazione superficiale. Un meccanismo analogo agisce in un fluido la cui superficie è esposta all'aria. La fotografia è di H. Linde dell'Istituto centrale di chimica fisica di Berlino. Le particelle plastiche sospese nei liquidi visualizzano il flusso.

una superficie piana. Così le interazioni della gravitazione e della tensione superficiale sono delicate e complesse. Una teoria che comprende sia le forze di spinta sia quelle di tensione superficiale è stata formulata nel 1964 da D. A. Nield dell'Università di Auckland.

Il predominio della tensione superficiale nella convezione di Bénard è stato ora accertato senza alcun dubbio. Una prova indiscutibile è che le caratteristiche celle esagonali di convezione appaiono anche quando lo strato fluido viene riscaldato da sopra anziché da sotto. In tali circostanze il gradiente di densità si oppone al flusso convettivo, cosicché le forze risultanti dalla tensione superficiale lo devono superare. Il flusso convettivo attribuito alla tensione superficiale è stato osservato anche in esperimenti condotti da due missioni spaziali Apollo, nelle quali la gravità e la spinta verso l'alto erano trascurabili.

La teoria di Rayleigh e le teorie basate su di essa forniscono approssimativamente le condizioni necessarie per l'innesco della convezione. Cosa accade una volta iniziato il flusso? Nella descrizione di tali moti convettivi completamente innescati le teorie di questo genere sono molto meno soddisfacenti.

Nella teoria di Rayleigh la velocità del flusso è data da una funzione esponenziale: la velocità è proporzionale al numero di Eulero, e (che ha un valore numerico di circa 2,7) elevato a una certa potenza. L'esponente è uguale al tempo, t (misurato in secondi, per esempio, a partire da un certo istante di riferimento) moltiplicato per un coefficiente, λ, determinato dal numero di Rayleigh. La velocità è perciò proporzionale a $e^{\lambda t}$. Le previsioni della teoria possono essere catalogate osservando l'evoluzione di questa espressione per differenti valori di λ.

Se il numero di Rayleigh è minore del numero critico di Rayleigh, λ è negativo. Con il passar del tempo, perciò, la velocità è data da potenze negative di e progressivamente crescenti; in tali condizioni il valore dell'esponenziale tende a zero al crescere di t. In altre parole, la velocità cade a zero e tutti i moti casuali del fluido vengono smorzati. Quando il numero di Rayleigh coincide esattamente con il valore ciritico, λ è nullo, e quindi l'esponente λt è sempre nullo. Dal momento che qualsiasi numero elevato a potenza nulla è uguale a 1, la perturbazione non è né smorzata né accelerata, ma mantiene indefinitamente il suo valore iniziale.

Entrambe le previsioni sono in accordo con l'intuizione e con l'analisi della stabilità dello strato fluido. Un valore negativo di λ corrisponde allo stato stabile di quiete, mentre il valore $\lambda = 0$ indica una stabilità marginale. L'interpretazione della teoria diventa però problematica quando il numero di Rayleigh è maggiore del valore critico e λ è positivo. Sono proprio le condizioni nelle quali si può instaurare un flusso convettivo.

Quando λ è positivo, la potenza a cui viene elevato e aumenta continuamente nel tempo e l'espressione riproduce il ben noto andamento di una crescita esponenziale. Se λ è uguale a $+1$ e la velocità iniziale è di un centimetro al secondo, dopo un secondo la velocità è aumentata a 2,7 centimetri al secondo e dopo due secondi ha raggiunto 7,4 centimetri al secondo. La velocità cresce illimitatamente, il che conduce subito a previsioni assurde; per esempio, la corrente di convezione prevista raggiunge la velocità della luce in meno di mezzo minuto.

La dipendenza della velocità del flusso dal valore di λ è stata qui esposta in forma semplificata. In generale λ è un numero complesso, un numero cioè che ha una parte reale e una parte immaginaria; quest'ultima comprende come fattore la radice quadrata di -1. Finora abbiamo considerato soltanto la variazione della parte reale di λ. Se la parte immaginaria non è nulla, può nascere un flusso oscillante, condizione detta di superstabilità. Tali correnti oscillanti sono state osservate nei fluidi reali e costituiscono una sottoclasse interessante dei fenomeni convettivi. Nel modello descritto dalla teoria di Rayleigh, la parte immaginaria di λ però scompare e si deve affrontare direttamente il problema della crescita esponenziale.

L'aumento di velocità di un flusso convettivo non può ovviamente continuare molto a lungo sulla curva esponenziale. Per tale ragione le previsioni della teoria di Rayleigh si possono considerare realistiche soltanto quando il numero di Rayleigh è prossimo al valore critico (in modo che λ sia piccolo) o soltanto per un breve intervallo di tempo successivo all'innesco della convezione (in modo che t sia piccolo). La causa fisica di tali limitazioni sta nelle ipotesi semplificative adottate nello sviluppo della teoria. In particolare abbiamo supposto che il gradiente termico fosse costante e quindi non influenzato dalla circolazione convettiva. Questa ipotesi è chiaramente contraria all'evidenza: quando un fluido caldo risale verso la parte superiore più fredda dello strato, la differenza di temperatura tra i contorni superiore e inferiore diminuisce. Diminuisce proporzionalmente la forza di spinta e il flusso pertanto si autolimita. Tuttavia, questo meccanismo di autolimitazione non compare nella matematica della teoria di Rayleigh, dove il gradiente termico resta costante indipendentemente dalla rapidità del flusso e la forza di spinta fornisce un impulso per un'accelerazione che continua indefinitamente.

Constatato il fallimento, può apparire sorprendente che la teoria di Rayleigh porti in tutti i casi a risultati accettabili. Può fare ciò soltanto perché comprende un'altra ipotesi, cioè che la porzione di fluido non riceve mai più di uno spostamento infinitesimo. Se è rispettata tale condizione, l'ipotesi di un gradiente termico costante è del tutto ragionevole. Perfino un movimento della porzione finito ma piccolo può provocare soltanto una piccola perturbazione della distribuzione di temperatura, in modo tale da lasciare ancora approssimativamente valide le previsioni della teoria. Tuttavia, nell'applicare la teoria di Rayleigh a un flusso completamente sviluppato, le ipotesi vengono violate e l'evoluzione del moto prevista finisce in un poco attendibile infinito.

Una teoria accettabile della convezione ben avviata deve in qualche modo scopri-

re il meccanismo di retroazione attraverso il quale il flusso stesso altera la forza che lo guida. Non si conosce alcun metodo pratico per risolvere esattamente il problema, ma esistono alcune approssimazioni che danno risultati migliori della teoria di Rayleigh. Nel 1937 il teorico sovietico L. D. Landau sviluppò la sua teoria in modo da descrivere certi tipi di transizioni di fase, quale l'innesco della magnetizzazione nel ferro ferromagnetico. Con V. L. Ginzburg, un altro fisico sovietico, egli la adattò più tardi alla descrizione della superconduttività nei metalli. Questi fenomeni hanno certi aspetti in comune con la convezione; in particolare, essi richiedono la descrizione simultanea delle fluttuazioni di struttura su molte scale di lunghezza. Quando si adatta la teoria di Landau al problema della convezione, essa comprende in prima approssimazione la teoria di Rayleigh.

Ciò che la teoria deve fornire è un'equazione di moto del fluido, un'equazione che esprime la velocità e l'accelerazione di una qualsiasi porzione di fluido per qualsiasi data combinazione di condizioni esterne imposte. L'equazione di moto si potrebbe costruire direttamente, ma il significato della teoria di Landau è più chiaro se si definisce invece una superficie di potenziale, dalla quale si può ricavare l'equazione di moto. (L'equazione di moto è semplicemente l'equazione che dà la pendenza della superficie di potenziale.)

La superficie di potenziale può essere immaginata come un panorama ondulato, nel quale l'altezza al di sopra o al di sotto di un qualche piano di riferimento rappresenta l'energia relativa del sistema fluido. La tendenza dell'energia ad assumere un valore minimo implica che il punto che rappresenta lo stato del sistema «rotoli verso il basso» appena possibile. Un asse sul piano di riferimento definisce la linea a velocità nulla; spostamenti di questa linea verso destra o verso sinistra indicano una velocità crescente positiva (verso l'alto) o negativa (verso il basso) di qualche porzione del fluido. La posizione sull'asse a velocità nulla è data dalla differenza tra il valore effettivo del numero di Rayleigh e il valore critico; tale differenza si può indicare con ΔR. Il valore di ΔR e della velocità V indicano insieme un punto sul piano di riferimento; l'altezza della superficie di potenziale in tal punto è l'energia del sistema in tale stato.

È necessario discutere in dettaglio l'equazione che descrive la topografia della superficie di potenziale. L'equazione si può scrivere come la somma di una serie infinita di termini, ciascuno dei quali contiene una potenza sempre più alta della velocità del fluido. Il primo termine è quadratico: $-\frac{1}{2}\Delta RV^2$. Nel termine successivo la velocità compare come V^3, nel successivo ancora come V^4 e così via. Ognuna delle potenze crescenti della velocità deve essere preceduta da un coefficiente che misura il suo contributo alla forma della superficie.

Come ci si potrebbe aspettare, è scomodo calcolare la somma dell'intera serie infinita; anche se disponessimo di metodi formali per farlo, non si conoscono i coefficienti di tutti i termini. Però, in generale si prevede che i coefficienti diventino più piccoli al crescere della potenza alla quale è elevata V. C'è quindi qualche speranza di ricavare una precisione ragionevole da una serie troncata, cioè una serie nella quale si trascurano tutti i termini al di là di una certa potenza di V. Se la velocità non è troppo grande, i contributi di tali termini di grado più elevato dovrebbero essere piccoli; in particolare, se V è minore di 1 in un certo sistema naturale di unità, le potenze più alte di V convergeranno a zero e l'approssimazione sarà buona.

La teoria di Rayleigh può essere vista come un tale troncamento della serie infinita, che tiene valido quindi soltanto il primo termine, cioè $-\frac{1}{2}\Delta RV^2$. La superficie descritta da questa espressione quadratica ha due lobi, ognuno dei quali con una sezione trasversale parabolica; un lato è concavo verso l'alto, mentre l'altro è concavo verso il basso. Dall'equazione è ovvio che quando V è nulla (cioè, in tutti i punti dell'asse a velocità nulla), l'energia relativa è nulla. Se il numero di Rayleigh è inferiore al valore critico, e quindi ΔR è trascurabile, la superficie si incurva verso l'alto ai due lati dell'asse e l'energia aumenta tutte le volte che la velocità è maggiore di zero. In altri termini, lo stato di quiete è un minimo della superficie di potenziale, uno stato di equilibrio stabile. Quando il numero di Rayleigh supera il valore critico, e quindi ΔR è positivo, la situazione è esattamente opposta: la superficie cade ai lati dell'asse a velocità nulla, che ora è una linea di equilibrio instabile e di energia massima.

Queste proprietà della superficie sono proprio quelle previste dall'analisi elementare di stabilità esposta sopra e comprendono sia i pregi sia i difetti della teoria di Rayleigh. Nelle immediate vicinanze dell'origine, dove tanto ΔR quanto V sono piccoli, si può dedurre correttamente il comportamento del sistema dalla variazione di curvatura della superficie. Quando ΔR è di poco negativo il fluido si riporta in quiete dopo una qualsiasi piccola perturbazione. Se ΔR è di poco positivo la perturbazione viene amplificata e inizia un flusso convettivo. Quando il numero di Rayleigh ha esattamente il valore ciritico, cioè ΔR è nullo, la superficie è piatta e una fluttuazione casuale di velocità non viene né smorzata né amplificata. Invece, per valori più alti di ΔR e di V, ci si ritrova di fronte a un problema che ormai dovrebbe esserci familiare. La superficie di potenziale pende verso meno infinito con il risultato di far aumentare illimitatamente la velocità.

La superficie di potenziale associata alla teoria di Rayleigh definisce l'energia relativa del fluido per qualsiasi combinazione del numero di Rayleigh e della velocità. La superficie è descritta da un'equazione quadratica nella quale il coefficiente, ΔR, è la differenza tra il valore effettivo e il valore critico del numero di Rayleigh. Fin quando ΔR è negativo, l'asse a velocità nulla rappresenta lo stato di minima energia e può essere mantenuto un flusso convettivo soltanto a condizione di aumentare l'energia. Quando ΔR è positivo, la pendenza della superficie viene invertita e l'asse a velocità nulla definisce uno stato di equilibrio instabile; il fluido può diminuire la propria energia instaurando una circolazione convettiva. Il difetto maggiore della teoria di Rayleigh è che, una volta iniziato il flusso, la velocità continua ad aumentare indefinitamente, previsione non realistica.

Nella teoria di Landau a questo difetto si può porre rimedio mantenendo altri termini della serie infinita. In realtà si ottengono buoni miglioramenti con la sola aggiunta di pochi termini. Per l'esperimento ipotetico dal quale è stata ricavata la teoria di Rayleigh la scelta dei termini è limitata da considerazioni di simmetria. In tale esperimento la natura della circolazione resterebbe immutata se l'apparecchiatura venisse riscaldata dal sopra anziché dal sotto, o se tutte le velocità invertissero direzione. Questa invarianza implica che la superficie di potenziale sia simmetrica rispetto all'asse a velocità nulla; una siffatta superficie simmetrica è descritta da un'equazione che comprende solo potenze pari della velocità (quali V^2, V^4 e così via). Quando l'esponente è un numero pari, V e $-V$ elevati alla stessa potenza danno lo stesso risultato, mentre i risultati sono di segno opposto quando l'esponente è un numero dispari. Quindi tutti i termini in cui V è elevato a una potenza dispari (quali V, V^3 e così via) devono avere coefficienti nulli.

Interessanti risultati si possono avere aggiungendo un altro termine nel potenziale di Landau oltre a quello quadratico, cioè il termine di quarto grado $\frac{1}{4}V^4$. La topografia della superficie di potenziale si determina in tal caso calcolando l'espressione $-\frac{1}{2}\Delta R\ V^2 + \frac{1}{4}V^4$. Quando ΔR è negativo la superficie assomiglia molto alla più semplice superficie quadratica, pur se l'energia aumenta più rapidamente quando la velocità si allontana da zero. Quando ΔR è positivo, la forma della superficie è invece alterata in modo significativo. L'energia diminuisce su entrambi i lati dell'asse a velocità nulla, ma la diminuzione non continua indefinitamente. L'energia raggiunge invece un valore minimo per poi risalire quando la velocità aumenta ancora. Sia il valore del minimo di energia sia la velocità alla quale si raggiunge il minimo aumentano con ΔR.

La teoria di Landau porta a una superficie di potenziale più attendibile aggiungendo alcuni termini addizionali all'equazione che definisce la superficie. Nel caso più semplice qui illustrato, viene aggiunto un termine di quarto grado al termine quadratico della teoria di Rayleigh. Lo stato stazionario di quiete diventa di nuovo instabile quando ΔR diventa maggiore di zero, ma la velocità non continua a crescere indefinitamente. Appare invece un minimo nella superficie di potenziale per una velocità finita. La teoria prevede un nuovo equilibrio alla velocità che rende minima l'energia totale; essa è stata avanzata nel 1937 da L. D. Landau per descrivere certe transizioni di fase in materiali magnetici ed è stata recentemente adattata al problema della convezione.

Per un limitato intervallo di numeri di Rayleigh e di velocità questa versione relativamente semplice della teoria di Landau può fornire previsioni realistiche. Come in precedenza, quando il numero di Rayleigh è minore di quello critico, viene smorzata qualsiasi fluttuazione casuale di velocità e lo stato stazionario è quello di minima energia e di equilibrio stabile. Per valori del numero di Rayleigh superiori a quello critico una perturbazione cresce con maggiore rapidità, ma la crescita non continua illimitatamente. Quando il fluido raggiunge una certa velocità finita, determinata dal valore di ΔR e corrispondente al minimo della superficie di potenziale, si ottiene un nuovo equilibrio stabile. Per qualsiasi valore di velocità diverso da questo si ha smorzamento.

La teoria di Landau che comprende termini quadratici e di quarto grado evita alcuni dei più spettacolari fallimenti della teoria di Rayleigh, ma è ancora un'approssimazione ed è valida soltanto se la velocità non è troppo grande. Quando V è grande, le potenze di V elevate forniscono un contributo significativo anche se sono precedute da un coefficiente piccolo; per questo motivo una teoria che trascura tutti i termini di potenza più alta non può rappresentare con precisione la forma della superficie di potenziale lontano dall'asse a velocità nulla. Inoltre, in molti sistemi convettivi, una direzione di flusso è privilegiata sull'altra, in modo da far perdere la simmetria della superficie di potenziale e nell'equazione devono essere inserite le potenze dispari di V.

Sia la teoria di Rayleigh sia quella di Landau sono ricavate da esperimenti ipotetici nei quali è considerato costante il maggior numero di proprietà del fluido. I fluidi reali sono raramente così semplici e le interazioni delle varie proprietà possono diventare assai complesse. Per esempio, nel modello si era supposto che solo la densità variasse in funzione della temperatura. In realtà anche la viscosità e la diffusività termica variano con la temperatura nella maggior parte dei fluidi. Dal momento che queste grandezze entrano nella definizione del numero di Rayleigh, una loro variazione può avere un'influenza importante sull'innesco di un flusso convettivo e sulla sua conseguente evoluzione. Si era anche supposto che il fluido fosse incomprimibile; dato che molti fluidi reali sono comprimibili, la pressione diventa una variabile significativa, che influenza a sua volta la densità e molte altre proprietà. La temperatura e la viscosità sono correlate da una relazione di notevole complessità. In generale la viscosità diminuisce all'aumentare della temperatura, ma nello stesso tempo l'energia dissipata dalla resistenza viscosa appare sotto forma di calore e fa quindi aumentare la temperatura.

Una teoria che tenga conto esplicitamente di tutte le interazioni conosciute delle proprietà del fluido sarebbe troppo scomoda per essere pratica. Nella descrizione di un sistema di un fluido reale più che un modello matematico ciò che si deve cercare è un valido compromesso tra la complessità del fluido e la complessità della teoria. La natura del compromesso necessario si può illustrare con alcuni esempi di convezione nel mondo esterno al laboratorio.

Nell'atmosfera terrestre si osserva la convezione su diverse scale di lunghezza. Il gradiente termico tra i tropici e i poli induce una circolazione globale che può essere decomposta in almeno tre grandi celle convettive in ogni emisfero. Le di-

storsioni di questi modelli causate dalla rotazione della Terra danno origine ai venti alisei dei tropici e ai venti occidentali prevalenti delle zone temperate. Il riscaldamento locale dell'atmosfera nei pressi della superficie terrestre dà origine a flussi convettivi su scala minore, tra i quali quelli della maggior parte delle tempeste. I cumuli, che si formano quando l'aria calda sale e si raffredda e diventa perciò sovrassatura di umidità, caratterizzano spesso lo sconvolgimento convettivo dell'atmosfera.

Un'analisi teorica della convezione atmosferica deve tener conto della grande comprimibilità dell'aria, che dà origine a un gradiente di densità anche quando la temperatura è costante con la quota. Una descrizione accurata della circolazione atmosferica dovrebbe anche comprendere il riscaldamento per compressione dell'aria quando essa scende in una regione di maggiore pressione. Anche la viscosità e altre proprietà dell'aria variano con la pressione e la temperatura e la presenza di vapore acqueo, che emette calore quando condensa, aggiunge un ulteriore livello di complessità. Le nuvole formatesi come risultato di una circolazione convettiva sono anch'esse instabili a un ulteriore moto convettivo: la nuvola si raffredda nella parte superiore per perdita di calore verso lo spazio e si riscalda nella parte inferiore per la radiazione assorbita dal suolo. Se l'entità di tali effetti è sufficientemente grande, all'interno della nuvola può instaurarsi una cella di convezione.

Nonostante queste complicazioni il moto convettivo nell'atmosfera manifesta spesso gli stessi modelli fondamentali osservati in più semplici esperimenti di laboratorio. Le formazioni rettilinee di nubi chiamate strie o code di cavallo sono prodotte da celle del tipo a rullo; le fotografie dei satelliti rivelano occasionalmente schieramenti di celle poligonali che si estendono per migliaia di chilometri quadrati. Tuttavia i risultati degli esperimenti di laboratorio non si possono semplicemente estrapolare alla scala atmosferica. Le celle di convezione in laboratorio sono sempre larghe quanto alte, mentre le celle atmosferiche sono molto più larghe, di un fattore fino a 50. Inoltre il verso di circolazione in esperimenti su piccola scala è sempre lo stesso (con i gas il flusso è diretto verso il basso nel centro di ogni cella), mentre nell'atmosfera si osservano entrambe le direzioni di flusso.

Anche la convezione negli oceani copre un ampio intervallo di scale, da un paio di metri fino alle dimensioni degli stessi bacini oceanici. I più semplici di tali flussi hanno una spiegazione immediata. Dato che alcune lunghezze d'onda della radiazione solare penetrano negli oceani per alcune decine di metri, l'acqua viene riscaldata fino a una notevole profondità. Il raffreddamento, del resto, è dovuto quasi interamente all'evaporazione e alla perdita di calore per conduzione e per irraggiamento verso l'atmosfera, processi che sono praticamente limitati alla superficie. Il calore viene quindi introdotto negli oceani a un livello inferiore a quello al quale viene asportato, e uno strato d'acqua di parecchi metri di spessore può diventare instabile alla convezione.

La comprimibilità dell'acqua di mare è piccola e può avere effetto sul moto convettivo soltanto alle maggiori profondità, ma c'è un'altra variabile che può avere un importante effetto: la salinità. La densità dell'acqua marina varia non soltanto con la temperatura ma anche con la concentrazione dei sali disciolti; la densità aumenta con la salinità. Il risultato è che possono cooperare due importanti fattori per determinare un gradiente di densità. L'interazione di questi fattori può originare nuovi tipi di moto convettivo che non appaiono quando è presente solo un unico gradiente.

Se la temperatura è più alta sul fondo di uno strato mentre la salinità è maggiore nella parte superiore, entrambi i gradienti favoriscono la convezione. Quando i gradienti di temperatura e di salinità agiscono in direzioni opposte, entrano in gioco effetti più delicati. Se l'acqua calda salata si trova al di sopra dell'acqua fredda dolce il gradiente di temperatura favorisce la stabilità, mentre il gradiente di salinità la ostacola. Anche se i due gradienti opposti cooperano per produrre una densità uniforme, la convezione può a volte instaurarsi per ragioni che hanno a che fare con effetti dissipativi che agiscono in modo diverso sui due gradienti. Il gradiente di temperatura viene dissipato principalmente per diffusione del calore, mentre il gradiente di salinità viene dissipato principalmente dalla diffusione molecolare delle molecole di sale e d'acqua. La diffusione del calore è molto più rapida, spesso di un fattore 100. Inizialmente la temperatura e la salinità dei due strati potrebbero essere regolate per conferire a essi densità esattamente uguali. Se una porzione del fluido caldo salato viene poi spostata verso il basso nello strato freddo dolce, essa perderà calore per molto tempo prima che la diffusione molecolare possa ridurre la propria salinità in modo significativo. Il risultato è che essa diventerà più densa e il moto sarà amplificato.

La disposizione opposta, con il liquido freddo dolce posto al di sopra del fluido caldo salato può portare al fenomeno oscillatorio chiamato superstabilità. Una porzione di acqua salina calda salendo di poco si raffredda, ma mantiene la propria concentrazione salina. Il risultato è che essa diventa più densa di quanto lo fosse inizialmente e ridiscende verso lo strato inferiore. In realtà, essa può superare la propria posizione originale continuando a oscillare attorno a essa. Le oscillazioni possono crescere o venire smorzate, a seconda dei valori dei due gradienti.

Uno dei sistemi convettivi più complessi è quello che funziona apparentemente nel mantello terrestre, creando una catena di dorsali sul fondo del mare e facendo migrare i continenti sulla superficie terrestre. Il calore che induce la circolazione non viene liberato sul contorno, ma nel volume del materiale, principalmente per effetto del decadimento di elementi radioattivi. In tali circostanze si forma un gradiente termico perché il calore viene perso dal sistema soltanto alla superficie, in modo che la temperatura aumenta con la profondità. Non c'è dubbio che il gradiente sia abbastanza grande da indurre la convezione, ma le proprietà del sistema sono così complicate e il mantello è talmente inaccessibile alle misure che la forma e le dimensioni del modello di convezione sono estremamente incerte. La viscosità aumenta bruscamente con la profondità e a una certa quota della zona di convezione il materiale subisce evidentemente una transizione da una fase cristallina a un'altra.

Su una scala molto più piccola si può osservare un interessante processo convettivo di considerevole complessità in una pellicola di vernice o di lacca che si sta asciugando. In questo caso la forza responsabile non è la spinta verso l'alto, ma la tensione superficiale, come negli esperimenti di Bénard. Il meccanismo responsabile in ultima analisi del flusso è l'evaporazione del solvente dalla superficie libera della pellicola. Se qualche perturbazione fa aumentare la velocità di evaporazione in una regione, tale regione si raffredda con conseguente aumento della sua tensione superficiale. Inoltre la tensione superficiale propria dei pigmenti o di altre grandi molecole della pellicola è solitamente maggiore della tensione del solvente, e quindi una mancanza di solvente fa aumentare la tensione superficiale indipendentemente dalla temperatura. Il liquido viene risucchiato attraverso la superficie verso zone di elevata tensione superficiale, dove esso scende fino alla base della pellicola e riprende il ciclo. Però, quando la concentrazione del solvente è diminuita, la viscosità aumenta e il numero di Marangoni finisce con lo scendere al di sotto del valore critico. A questo punto la convezione cessa.

Le celle di convezione nelle pellicole di vernice hanno spesso una forma esagonale, o almeno una forma poligonale che si avvicina a quella ideale di esagoni regolari. Il flusso può provocare l'«accumulo» dei pigmenti che si manifesta nella pellicola asciutta come una colorazione irregolare. In alcuni casi il modello tridimensionale delle celle di convezione rimane congelato nella pellicola asciutta. Va notato che il fenomeno non è sempre indesiderabile: la pittura a «martello» acquista il suo aspetto in questo modo.

La generalità del concetto di convezione è suggerita da questi differenti esempi: lo sconvolgimento spontaneo dell'atmosfera e degli oceani terrestri, e la circolazione della vernice in una pellicola spessa pochi decimi di millimetro. Le teorie che descrivono questi moti fluidi richiedono molte ipotesi semplificative se devono essere di impiego pratico e anche così sono ben lontane dall'essere semplici. È perciò importante sottolineare che queste teorie correlate, dominate da pochi numeri adimensionali, possono render conto di fenomeni in scale tanto differenti.

Il DNA a superelica

In molte forme di DNA, la doppia elica si dispone a sua volta a elica. Lo avvolgimento a superelica, che ha importanti conseguenze biologiche, può essere adeguatamente rappresentato con un semplice modello matematico

di William R. Bauer, F. H. C. Crick e James H. White

Il DNA, il materiale genetico primario della maggior parte degli organismi viventi, viene in genere rappresentato come una doppia elica, in cui due catene di nucleotidi complementari (le subunità la cui sequenza dà luogo al messaggio genetico) si avvolgono attorno a un asse rettilineo comune. Risulta oggi evidente, però, che quest'asse, spesso, non è rettilineo, ma curvo. In effetti, la doppia elica può avvolgersi nello spazio a formare una nuova elica di ordine superiore. Si parla, in questo caso, di superelica e sembra che una buona percentuale dei DNA noti presenti questo tipo di avvolgimento almeno in uno stadio del ciclo vitale. Una valutazione di questa caratteristica strutturale e delle sue conseguenze è, dunque, essenziale per capire completamente la biologia del DNA.

Interessando DNA in un'ampia gamma di dimensioni e di forme (compresi alcuni che non hanno un'organizzazione a doppia elica), l'avvolgimento a superelica assume aspetti diversi. Per esempio, nella cromatina (complesso di DNA con una proteina) degli organismi superiori, il DNA si avvolge attorno a un'anima proteica e forma una superelica solenoidale sinistrorsa. Ci si occuperà qui, però, in prevalenza di un altro tipo di avvolgimento in cui non è necessaria la presenza di un'anima proteica. Si tratta di quello del DNA circolare, molecola a doppia elica in cui ogni catena polinucleotidica forma un anello chiuso.

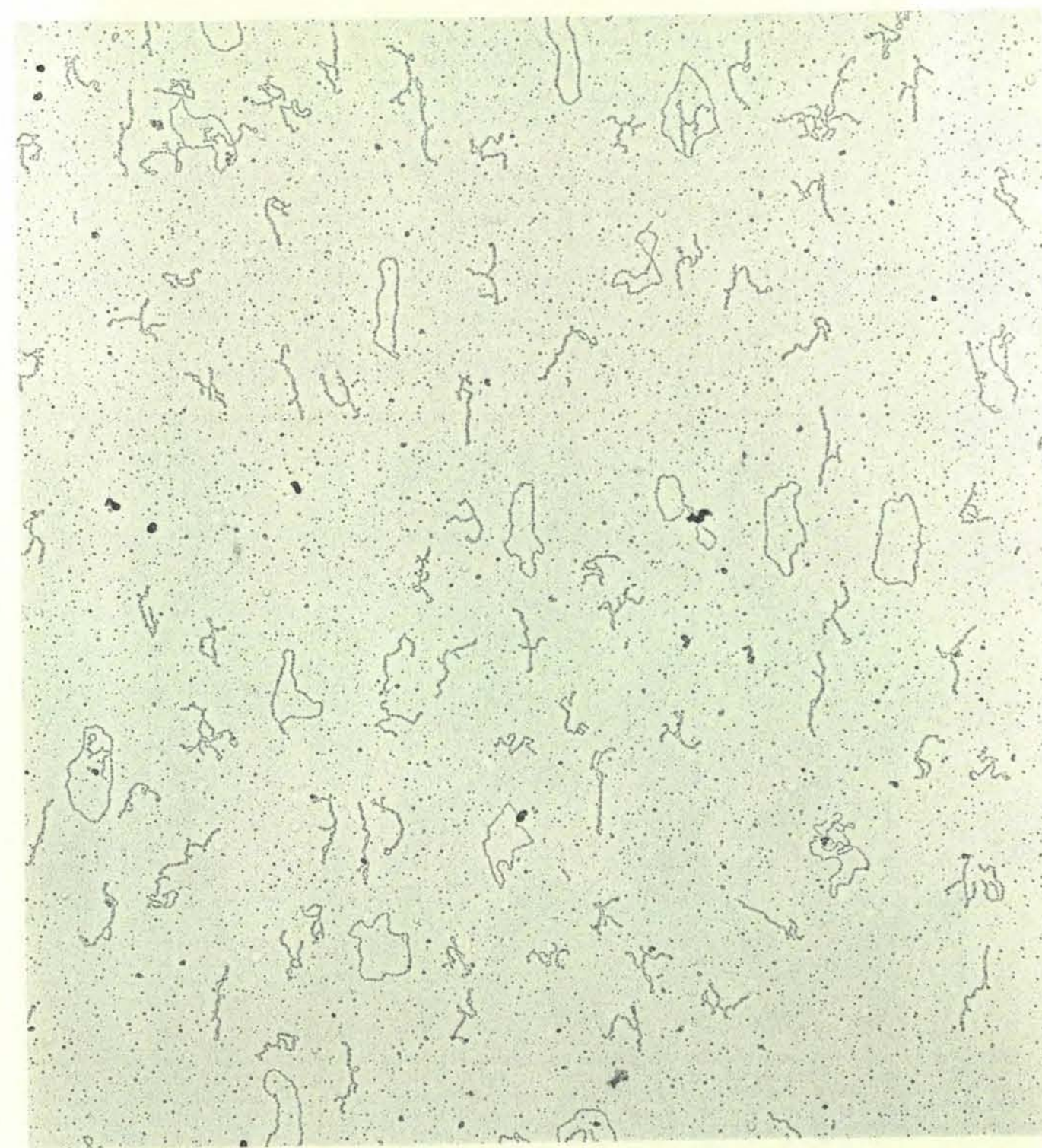
L'esistenza di questi anelli di DNA avvolti a elica è stata postulata inizialmente per spiegare una sorprendente scoperta sul DNA del virus del polioma, un piccolo virus che provoca tumori nel topo. Nel 1963 è stato riportato che, quando questo DNA viene sospeso in un solvente e centrifugato, esso si risolve in tre componenti, distinguibili per la velocità con cui si spostano nel solvente stesso. Ulteriori indagini hanno rivelato che le tre frazioni, contrassegnate con I, II e III, in ordine di velocità di sedimentazione decrescente, non differivano tra loro per peso molecolare. Pertanto doveva esserci, nella compattezza molecolare delle tre specie, qualche variazione, che poteva tener conto della loro diversa velocità. In altre parole, quelle molecole dovevano avere forme diverse.

Osservandole al microscopio elettronico, si è avuta la conferma che le cose stavano proprio in questi termini. Mentre le molecole del componente III erano chiaramente lineari, quelle dei componenti I e II mancavano di estremità. La rivelazione che questi due componenti erano circolari non servì, tuttavia, a risolvere il problema. Non c'era ancora alcun indizio per capire che cosa distingueva l'uno dall'altro. Allora, nel 1965, Jerome Vinograd del California Institute of Technology avanzò un'ingegnosa ipotesi, che risolse il mistero e introdusse il concetto di avvolgimento a superelica del DNA.

Vinograd suggerì che una molecola circolare chiusa di DNA (una molecola circolare in cui ambedue le catene polinucleotidiche fossero completamente integre) dovesse avere in genere un avvolgimento a elica meno compatto di un frammento lineare di DNA della stessa lunghezza. In altre parole, nelle molecole circolari del componente I vi sono meno giri d'elica di quelli delle molecole o del componente II o del componente III. Il componente I si comporta come se, prima che le estremità della doppia elica si congiungessero, fossero avvenuti, in direzione opposta a quella del normale avvolgimento a elica, un certo numero di giri di 360 gradi.

La differenza tra componente I e componente II del DNA del polioma sta quindi nel fatto che le molecole circolari del componente II presentano delle «intaccature»: in ogni molecola vi è almeno una tacca, o rottura, in corrispondenza di una delle due catene polinucleotidiche. Al contrario, le molecole circolari del componente I sono integre e hanno un minor numero di avvolgimenti. Dato che le forze che rendono stabile la doppia elica sono rilevanti, le molecole circolari chiuse resistono a questo avvolgimento a elica meno compatto e, per motivi che verranno discussi più avanti, per compensazione formano delle supereliche. Pertanto, le molecole circolari integre del componente I del DNA del virus del polioma sono più compatte delle molecole circolari con intaccature del componente II e, di conseguenza, hanno una velocità di sedimentazione più elevata.

L'avvolgimento a superelica è un fenomeno diffuso, che caratterizza molti DNA tra i più interessanti dal punto di vista medico e biologico. Una intera classe di virus tumorali a DNA, compresi i virus dei poliomi e quelli del papilloma umano, contengono questo DNA e il DNA dei mitocondri (gli organelli cellulari trasduttori di energia) delle cellule umane e di altre cellule animali forma delle supereliche. Si è anche trovato che le cellule animali contengono una vasta serie di piccolissime molecole di DNA a superelica, la cui funzione non è stata ancora determinata. È particolarmente sorprendente che la maggioranza dei più piccoli genomi (corredi di geni) noti rientri in questa categoria, ivi compresi i fattori genetici per la fertilità e per la resistenza ai farmaci. Inoltre, sembra che il fenomeno dell'integrazione, in cui un piccolo frammento di DNA viene inserito fisicamente in una molecola più grande di DNA, esiga che tale frammento sia avvolto a superelica. Quindi, i vettori, o veicoli, del DNA, nell'ingegneria genetica, sono di questo tipo. Forse il fatto più importante è che i virus tumorali a RNA, ivi compreso il virus cancerogeno di Epstein-Barr, utilizzano spesso l'enzima trascrittasi inversa per produrre copie di DNA, che poi si avvolgono a superelica nel corso dell'infezione. Si dovrebbe anche notare che le supereliche sono state riscontrate sia in molecole con soli 350 nucleotidi per catena polinucleotidica sia in molecole con anche 1 750 000 nucleotidi per catena.

Naturalmente, ogni proprietà fisica, chimica e biologica del DNA - trascrizione, comportamento idrodinamico, energetica, enzimologia, e altre ancora - è influenzata dall'essere il DNA una molecola circolare integra e dalle deformazioni associate con il suo avvolgimento a superelica. Il comprendere il meccanismo di quest'ultimo e le conseguenze di una simile caratteristica strutturale per il DNA pone, invece, problemi di considerevole complessità matematica. Fortunatamente, in questa impresa, due branche della matematica offrono un sostanziale aiuto: la topologia, che studia le proprietà delle strutture che non si modificano quando tali strutture si deformano, e la geometria differenziale, che applica i metodi del calcolo differenziale allo studio delle curve e delle superfici. Nelle pagine seguenti descriveremo innanzitutto un modello matematico di DNA circolare e, quindi, discuteremo le implicazioni del modello per il DNA reale. Considereremo anche alcuni metodi che sono stati elaborati per misurare l'avvolgimento a superelica e i suoi effetti.

In questa microfotografia elettronica si notano molte molecole di DNA avvolte a superelica. Ogni molecola è un anello di DNA a doppia elica. Quando gli anelli si aprono, uno dei filamenti della doppia elica risulta «intaccato», o rotto, per cui la molecola risulta decontratta, non avvolta a superelica. Laddove, invece, gli anelli non sono aperti, ma attorcigliati, essi non presentano intaccature. In queste molecole, il DNA ha, in generale, un numero di giri d'elica inferiore a quello di una molecola lineare della stessa lunghezza. Negli anelli senza intaccature, la tensione introdotta da questo minor avvolgimento è compensata da un avvolgimento a superelica. Le molecole qui rappresentate sono plasmidi estratti dal batterio *Escherichia coli* e separati da altre molecole di DNA di *E. coli* mediante un processo di centrifugazione. Sono ingrandite di circa 25 000 diametri. La microfotografia è stata scattata da Gary Cohen della State University of New York a Stony Brooks.

Si consideri, innanzitutto, la struttura fisica della doppia elica: è formata da due impalcature di sostegno zucchero-fosfato a cui si attaccano le basi nucleotidiche del DNA. Le basi sulle due impalcature che si fronteggiano sono appaiate, con circa 10 paia di basi per ogni giro d'elica. (Il numero esatto dipende dalla particolare configurazione della molecola di DNA.) Per studiare matematicamente l'avvolgimento a superelica, il metodo più conveniente è quello di costruire un modello in cui la struttura sia rappresentata da uno stretto nastro di spessore infinitesimale, che si attorciglia. Il sistema più semplice per costruire un simile modello è di specificare che i bordi del nastro seguano le impalcature zucchero-fosfato del DNA. Portando avanti l'operazione nel caso semplice di una molecola lineare a doppia elica, si dimostra però che questa scelta non è soddisfacente; l'asse del modello (cioè quella linea che è a metà strada tra i margini del nastro) si avvolge attorno all'asse rettilineo della doppia elica. Per i nostri scopi, è preferibile, invece, scegliere un modello il cui asse coincida con quello della doppia elica. In più, specifichiamo che il nastro deve sempre essere perpendicolare rispetto alle pseudodiadi, o assi di rotazione avvolti due volte, che sono distribuite lungo la doppia elica *(si veda l'illustrazione in basso).* (Una diade è associata con ogni coppia nucleotidica e l'altra con lo spazio tra coppie successive; questi assi sono tutti perpendicolari all'asse della doppia elica e la loro localizzazione è indipendente dalla sequenza delle basi nella molecola.)

Questo modello a nastro segue l'asse della doppia elica di DNA e si attorciglia alla stessa stregua di come si avvolgono le due catene della molecola attorno a quell'asse. Inoltre, dato che le sequenze di atomi nelle due catene polinucleotidiche decorrono in direzioni opposte, verranno assegnati ai bordi del nastro orientamenti opposti. (Non ha importanza quali direzioni vengano prescelte per i due bordi fintantoché decorrono in senso inverso.) Il modello può essere analizzato matematicamente in un certo numero di modi diversi, ma ci interesseremo qui soprattutto del rapporto che intercorre tra bordi del nastro orientati in direzioni opposte.

Quando le estremità di un nastro si congiungono, ogni bordo descrive una curva chiusa nello spazio tridimensionale e quando quel nastro rappresenta una molecola circolare di DNA, prima che le sue estremità si congiungano, forma un certo numero di giri di 360 gradi; in questo modo, le due curve descritte dai suoi bordi si uniscono e, in altre parole, risulta impossibile separarle senza «tagliarne» una. Se, in una coppia, ogni struttura circolare rappresenta una molecola unita con legami covalenti, come nel caso delle due catene polinucleotidiche della doppia elica, si dice che le due unità della coppia sono unite da un legame topologico. È questo uno speciale tipo di legame, nel senso che, pur non essendo alcuna parte di una molecola unita a qualsiasi altra parte dell'altra molecola con legami covalenti, cionondimeno è necessario rompere un legame covalente per poter separare le due molecole. (Il concetto di formazione di legami topologici è stato introdotto per la prima volta negli studi di chimica da Edel Wasserman e collaboratori dei Bell Laboratories.)

In termini matematici, il legame di due curve chiuse è una proprietà topologica: senza tener conto di come sono deformate (tirate, avvolte, ecc.), fintantoché né l'una né l'altra si rompe, rimarranno legate esattamente allo stesso modo. Sarà, quindi, utile assegnare un valore numerico che descriva il modo in cui la struttura circolare formata da un margine del nastro, che rappresenta una molecola di DNA, è legata alla struttura circolare formata dall'altro margine.

Sembra ragionevole definire un simile numero di legame Lg in modo che il suo valore sia 0 nel caso di una coppia di curve chiuse non unite tra loro, 1 per una curva che forma un «anello» attraversandone un'altra solo una volta, e così via. Tuttavia, poiché sono stati attribuiti gli orientamenti ai due bordi del nastro, si può formulare una definizione più utile del numero di legame, assegnando o un segno positivo o un segno negativo al numero che dipende dall'orientamento delle curve. Vi sono parecchi modi di calcolare un simile numero di legame per una particolare configurazione di curve e uno dei più convenienti consiste nell'esaminare tutti i punti di una proiezione, o rappresentazione bidimensionale, della configurazione in cui un pezzo di una delle curve incrocia un pezzo di un'altra curva *(si veda la figura della pagina a fronte).* A ciascuno di questi punti di intersezione si può assegnare un numero indice di + 1 o di – 1, in base alla direzione in cui il pezzo in alto deve essere ruotato per farlo coincidere con il pezzo in basso. Sommando i numeri indice di tutti i punti d'intersezione e dividendo la somma per due (il numero delle curve unite), si ottiene il numero di legame Lg.

Così, tale numero viene definito come un numero intero con segno + o –, che descrive una proprietà di due curve chiuse nello spazio. Per separare una coppia di curve senza tagliarle, il valore di Lg deve essere 0 (anche se l'inverso non è sempre

Il modello del DNA a superelica sfrutta due caratteristiche strutturali della doppia elica *(a sinistra):* l'asse *(linea spessa in colore),* attorno al quale le due impalcature zucchero-fosfato della molecola si avvolgono, e la serie di «pseudodiadi», assi definiti dalle basi nucleotidiche attaccate alle impalcature suddette. Le basi, la cui particolare sequenza costituisce il messaggio genetico, sono appaiate e a ogni coppia di basi è associata una diade, mentre un'altra diade è associata allo spazio tra coppie di basi successive. (Qui viene raffigurata soltanto una coppia di basi.) L'asse della diade *(linea sottile in colore),* definito da ogni coppia di basi, è la linea perpendicolare all'asse dell'elica attorno al quale l'una o l'altra delle due basi della coppia può essere fatta ruotare nella posizione della base contrapposta. Il modello costruito secondo queste caratteristiche *(a destra)* è un nastro stretto, di spessore infinitesimale, il cui asse, o linea centrale, segue l'asse della doppia elica e la cui superficie è sempre perpendicolare alle diadi, assi definiti dalle coppie di basi della doppia elica. Questo nastro si piega quando anche l'asse della molecola di DNA si piega e si attorciglia allorché le due catene polinucleotidiche del DNA si avvolgono l'una attorno all'altra.

Il DNA circolare chiuso è rappresentato da questo modello che consta di un nastro attorcigliato le cui estremità sono state congiunte *(a).* Dato che le sequenze degli atomi nelle due catene della doppia elica decorrono in direzioni opposte, è utile assegnare anche ai bordi del nastro direzioni opposte. Quando tali margini sono visti come curve chiuse orientate nelle tre dimensioni dello spazio *(b),* si trova che sono matematicamente unite; in altre parole, non c'è modo di separarle senza rompere l'una o l'altra. Questa relazione può essere descritta matematicamente da un numero di legame Lg, la cui entità esprime il numero di volte che una curva si lega attraversando l'altra; il suo segno dipende dal modo in cui le curve sono marcate. Si può calcolare questo valore esaminando una proiezione, o rappresentazione bidimensionale, delle due curve e, a ogni punto in cui una curva passa sopra all'altra (non quando una curva passa sopra a se stessa) si assegna un numero indice secondo la seguente regola: se è necessaria una rotazione in senso orario per spostare il pezzo in alto, in modo che coincida con il pezzo in basso, allora si assegna al punto di incrocio il valore di +1; se è necessaria, invece, una rotazione in senso antiorario, si assegna il valore di –1 *(c).* (È il contrario di ciò che normalmente si fa in matematica, ma così il numero di legame del DNA circolare chiuso destrorso è positivo.) Si calcola, quindi, il valore Lg sommando i numeri indice e dividendo per 2 (il numero delle curve). Il numero di legame ottenuto in questo modo è un numero intero relativo, uguale a 0 se le due curve non sono legate *(d),* a +1 o –1 se una curva si lega all'altra in un unico punto *(e),* a +2 o –2 se una curva si lega all'altra due volte *(f)* e così via. Il segno si modificherà se l'orientamento dell'una o dell'altra curva cambierà o se la coppia di curve viene vista in uno specchio *(g).* Il valore del numero di legame rimane lo stesso indipendentemente da come le due curve si deformano *(h)* e così, dato che il numero di legame di un nastro attorcigliato è uguale a quello delle impalcature zucchero-fosfato della molecola di DNA di cui rappresenta il modello, esso esprime un'importante vincolo alla possibile struttura a superelica del DNA.

vero). Se le curve in questione sono i bordi di un nastro con le estremità unite e con N avvolgimenti, il numero di legame che le contraddistingue sarà $+N$ o $-N$, secondo la direzione dell'avvolgimento. Inoltre, fintantoché le estremità del nastro tenderanno a rimanere unite, il numero suddetto rimarrà identico anche quando il nastro si deformerà. Si noti che, pur non essendo stati scelti i bordi del modello a nastro del DNA in modo che coincidessero con le impalcature di sostegno zucchero-fosfato della doppia elica, il numero di legame del nastro sarà esattamente quello delle impalcature.

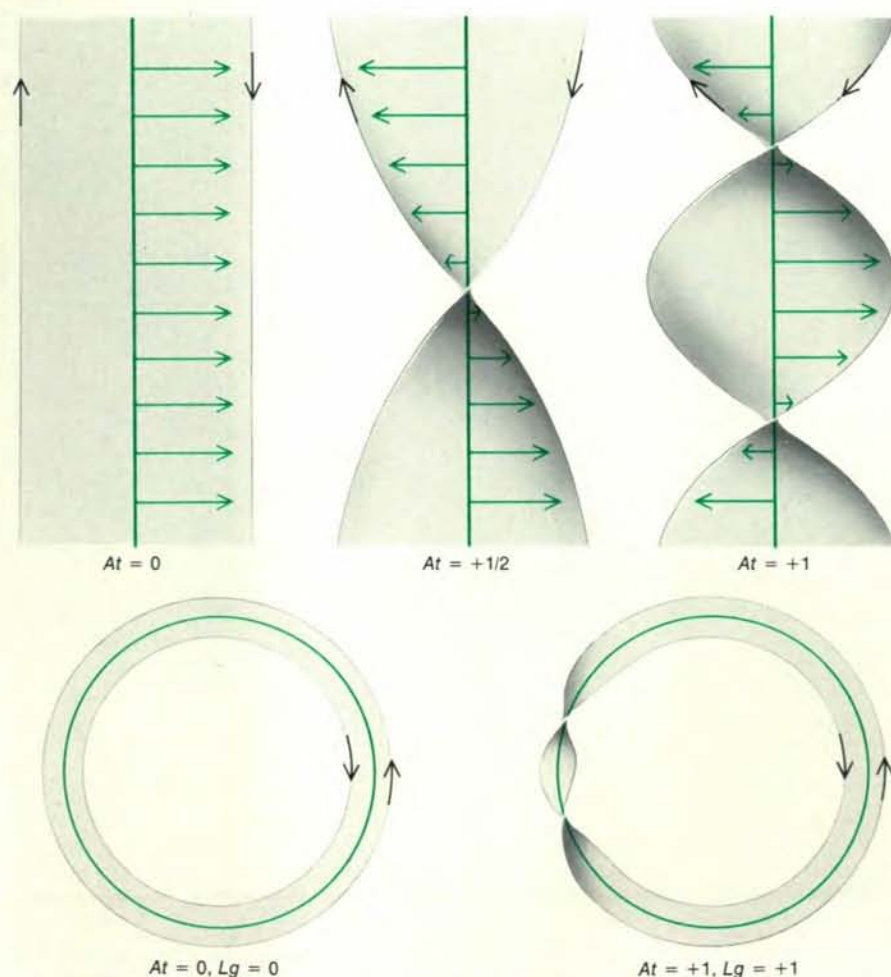
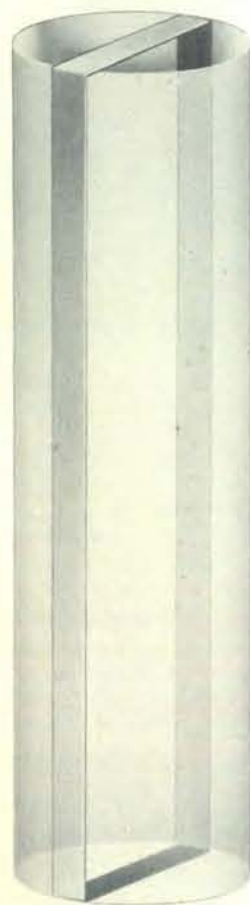
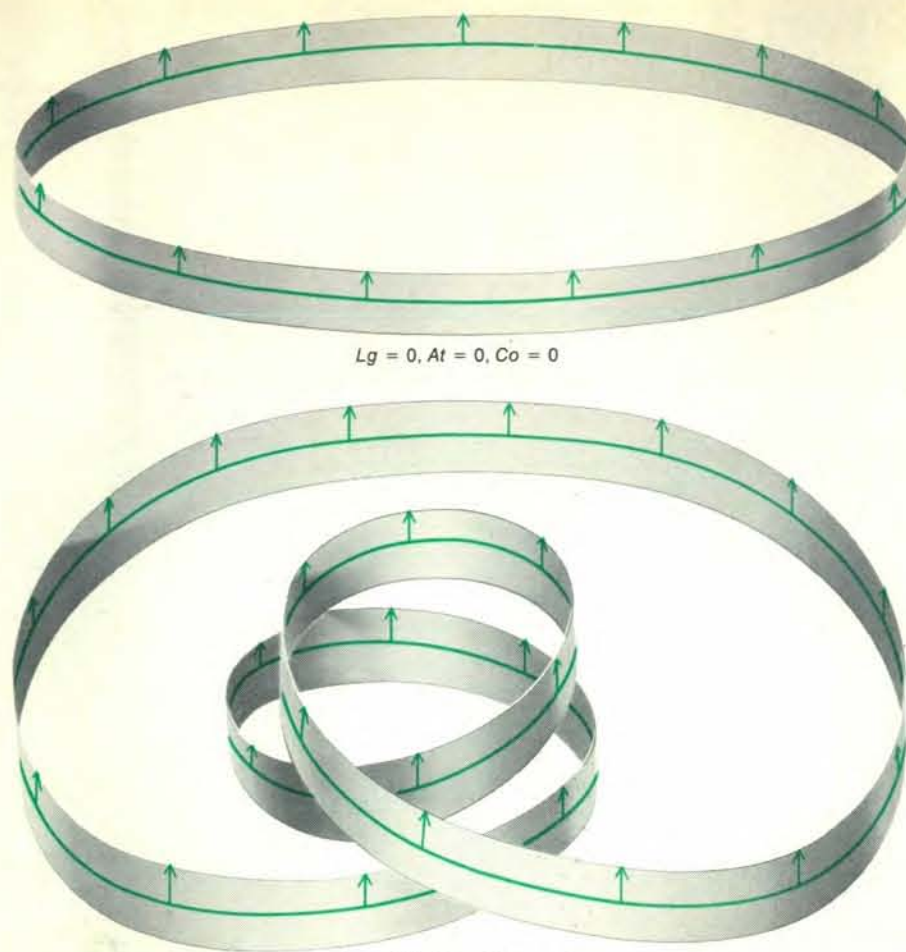
Pertanto, per una molecola di DNA circolare chiusa e decontratta, di una lunghezza pari a 5000 coppie di basi, con 10 coppie per ogni giro d'elica, il numero di legame sarà + 500. Per convenzione, tale numero è positivo perché la doppia elica di DNA è destrorsa. (Una molecola di DNA circolare chiusa è decontratta se il suo asse giace interamente in un piano. Come riporteremo, esistono molti modi in cui si riescono a far decontrarre molecole naturali avvolte a elica. Una molecola circolare rotta è decontratta allo stato naturale, ma dato che le sue due catene polinucleotidiche non sono entrambe intatte, rigorosamente parlando, non possiede un numero di legame, che è la grandezza che interessa maggiormente qui.)

Un altro modo per analizzare il modello a nastro del DNA è quello di osservare non il rapporto tra i suoi bordi, ma di osservare il modo in cui il nastro si attorciglia. Per un nastro, il cui asse segua una linea retta, l'idea di un valore numerico che esprima l'avvolgimento è intuitiva. Si adotterà qui la convenzione per cui un giro destrorso di 360 gradi ha un valore di + 1 e un giro sinistrorso, sempre di 360 gradi, un valore di – 1. La definizione di giro è meno chiara, invece, per un nastro il cui asse non sia rettilineo. Forse il miglior modo per capire questo concetto è immaginare una piccola freccia posta perpendicolarmente all'asse del nastro, puntata verso uno dei bordi *(si veda la figura in questa pagina)*. Spostandosi lungo il nastro che si attorciglia, essa ruota attorno all'asse e il giro che compie il nastro può essere definito come l'integrale della velocità angolare di rotazione della freccia rispetto alla lunghezza dell'arco formato dalla curva dell'asse.

Nel caso speciale in cui l'asse del nastro sia situato in un piano, si può misurare questo valore semplicemente come numero di rotazioni che la freccia compie attorno all'asse, muovendosi lungo il nastro. Per esempio, quando il nastro rappresenta come modello un frammento circolare chiuso di DNA, della lunghezza di 5000 coppie di basi e decontratto (cioè il cui asse giace in un piano), la freccia compie una rotazione completa per ogni giro della doppia elica e così l'avvolgimento totale *At* è uguale a + 500, con la comparsa, di nuovo, del segno + perché la doppia elica è destrorsa.

Dunque, per un frammento circolare decontratto di DNA della lunghezza di 5000 coppie di basi, sia il numero di legame sia l'avvolgimento risultano pari a + 500. Da questo esempio, si potrebbe allora presumere che il primo non è che un altro modo di esprimere il secondo, ma in realtà non è così. In effetti, è particolarmente importante capire la distinzione tra queste due grandezze. Per cominciare, il legame è una proprietà topologica, mentre l'avvolgimento è una proprietà geometrica: se un nastro si deforma, il suo avvolgimento si modifica. Inoltre, per calcolare il numero di legame (che è sempre un numero intero), il nastro va considerato come un tutto. D'altra parte, si può considerare l'avvolgimento (che potrebbe anche non essere rappresentato da un numero intero) da un punto di vista locale e i suoi valori per la singole sezioni possono poi essere sommati, così da ottenere il totale per tutto il nastro.

È importante notare che questo modello a nastro non si avvicina poi tanto alla rappresentazione delle proprietà meccaniche particolareggiate del DNA, non ultimo perché si ammette che esso abbia uno spessore infinitesimale. Un nastro reale (con uno spessore finito) sottoposto alle manipolazioni che interessano e che lo trasformano in un'elica subirebbe numerose deformazioni locali, che non sono prese in considerazione qui. Il modello a nastro, invece, è stato ideato per mettere in rilievo il numero di legame del DNA, il quale, essendo quantizzato, non è alterato da queste piccole deformazioni.

Il realizzare che il legame e l'avvolgimento sono due proprietà distinte fa sorgere un altro interrogativo: c'è forse un significato geometrico nella differenza tra queste proprietà, cioè nella differenza tra numero di legame di un nastro e il suo avvolgimento totale? Quasi casualmente, all'incirca nello stesso periodo in cui i biochimici stavano studiando per la prima volta il DNA circolare, i matematici, del tutto indipendentemente da essi, stavano

All'avvolgimento di un nastro si può assegnare un valore numerico ponendo una piccola freccia sul nastro stesso, perpendicolarmente al suo asse e orientato verso uno dei suoi bordi. A mano a mano che la freccia si sposta lungo il nastro ruota attorno all'asse e l'entità dell'avvolgimento totale *At* può essere definita come l'integrale della velocità angolare di questa rotazione rispetto alla lunghezza dell'arco della curva descritta dall'asse. Che questa quantità sia positiva o negativa dipende dal fatto che la rotazione della freccia attorno all'asse sia rispettivamente destrorsa o sinistrorsa. Quando l'asse del nastro giace completamente in un piano, come in questo disegno, l'avvolgimento totale si calcola facilmente, poiché è uguale al numero delle rotazioni che la freccia compie attorno all'asse. L'avvolgimento può essere calcolato separatamente per differenti parti del nastro e i valori parziali si possono poi sommare per poter ottenere il valore totale.

analizzando la formazione di legami e l'avvolgimento nei nastri. Nel 1968, uno di noi (White) ha dimostrato che il numero di legame di un nastro e il suo avvolgimento complessivo differiscono di una quantità che dipende esclusivamente dalla curva che compie l'asse del nastro (quantità ben nota ai matematici come integrale di Gauss della curva dell'asse). In altre parole, si ammette che gli assi di due nastri vicini seguano la stessa curvatura nelle tre dimensioni dello spazio; in tal caso, anche se i due nastri si girano e si avvolgono in modi totalmente diversi, i valori del loro numero di legame e dell'avvolgimento complessivo differiranno esattamente della stessa entità.

All'incirca nello stesso periodo, Vinograd, non essendo a conoscenza di questo risultato, chiese a F. Brock Fuller, un matematico del Cal Tech, di affrontare la matematica dell'avvolgimento a superelica. Fuller, lavorando sul rapporto tra legame e avvolgimento, suggerì il pittoresco termine di «numero di contorcimento» per indicare l'entità della differenza tra le due proprietà sopra citate. Così, per un nastro chiuso nello spazio tridimensionale, tale numero (*Co*) è uguale alla differenza tra numero di legame *Lg* e avvolgimento totale *At*: $Co = Lg - At$. Il numero di contorcimento di un nastro è una quantità rilevante, il cui valore generalmente cambia se l'asse del nastro si deforma nello spazio. Pertanto il contorcimento, come l'avvolgimento, non è una proprietà topologica del nastro, ma geometrica.

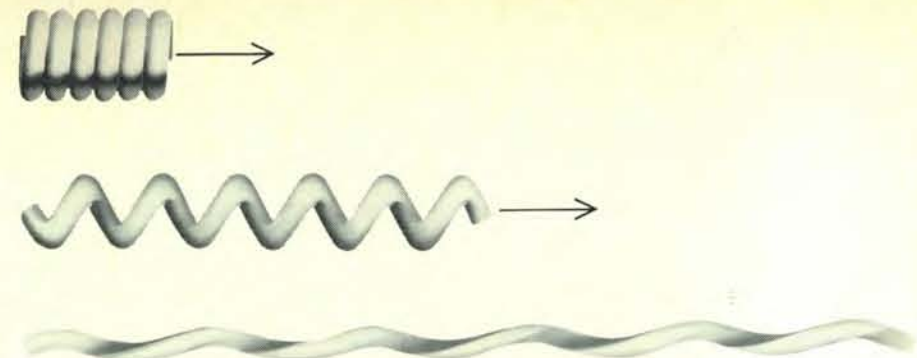
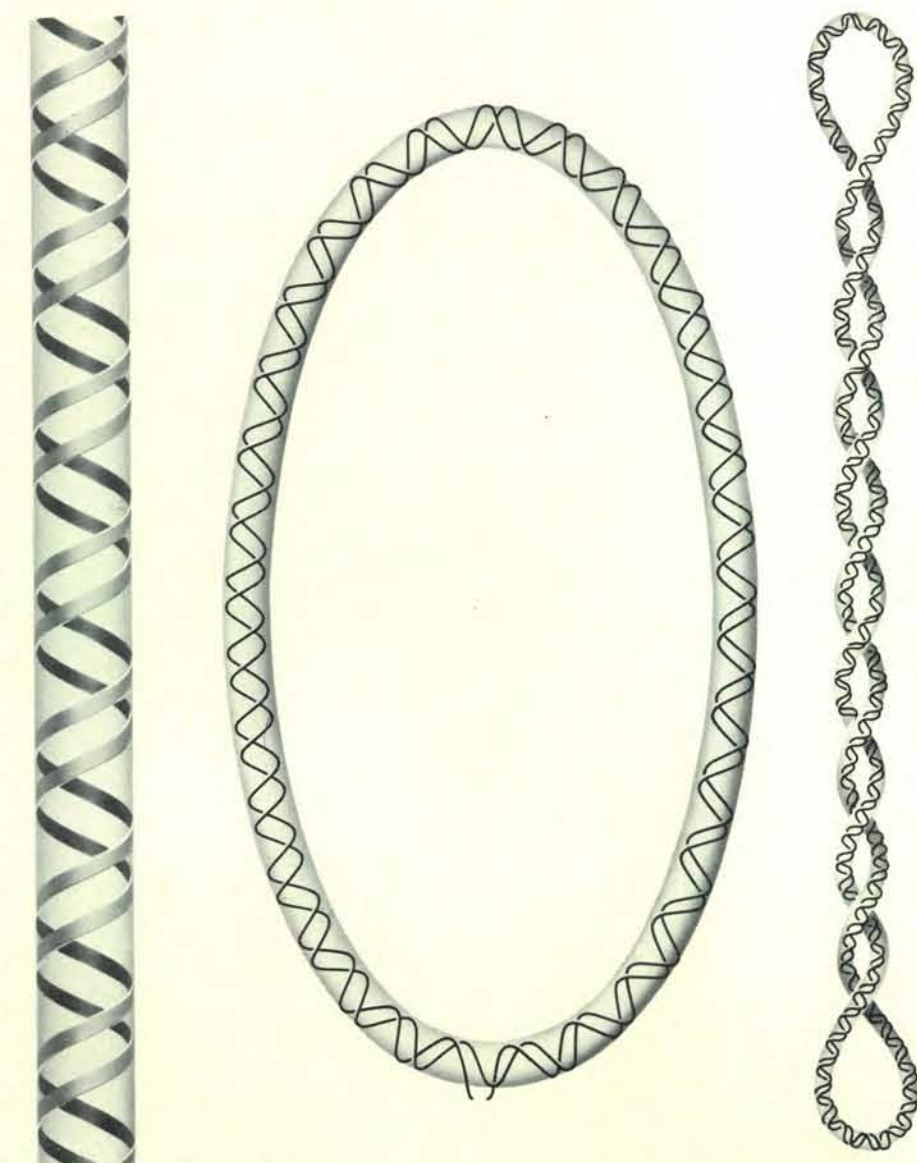
Il numero di contorcimento si può ricavare calcolando l'integrale di Gauss, ma in generale è molto più facile ottenerlo valutando il numero di legame e l'avvolgimento totale del nastro in questione e quindi calcolando la differenza. Solo in certi casi speciali, è conveniente calcolarlo direttamente. Per esempio, se l'asse di un nastro giace interamente in un piano o

Il numero di legame e l'avvolgimento totale di un nastro sono uguali quando l'asse del nastro giace completamente in un piano, come appare qui, in alto. Che queste due quantità non debbano necessariamente avere lo stesso valore è indicato dalla configurazione al centro, in cui il numero di legame *Lg* è un numero intero positivo mentre l'avvolgimento totale *At* si avvicina a zero. Il numero di legame e l'avvolgimento totale di un nastro differiscono di una quantità, chiamata numero di contorcimento, *Co*. Cioè: $Co = Lg - At$. Il numero di contorcimento dipende esclusivamente dalla curva che compie l'asse del nastro. Nel nastro che si avvolge attorno al cilindro illustrato in basso a sinistra, *Lg*, *At* e *Co* sono tutti zero. La rotazione della sommità del cilindro in senso antiorario attorno al suo asse centrale per due giri completi genera il nastro a elica destrorsa con avvolgimenti che si intersecano, che compare in basso a destra e in cui *Lg* risulta chiaramente immutato, mentre (trascurando gli effetti finali) *At* è uguale a $+ 4 \text{ sen } \alpha$ e *Co* è uguale a $- 4 \text{ sen } \alpha$. (Nel primo caso l'avvolgimento totale *At* è stato reso pari a zero, per cui la configurazione finale è un modello semplificato del DNA a superelica, omettendo l'avvolgimento intrinseco che risulta dai giri della doppia elica.)

interamente sulla superficie di una sfera, si può dimostrare che il numero di contorcimento è zero. Sostituendo questo valore nell'equazione $Co = Lg - At$ si ottiene $Lg = At$, che spiega perché, nell'esempio della molecola circolare chiusa decontratta di DNA, si sia trovato che sia *At* sia *Lg* erano + 500. Si consideri ora che cosa succede se l'asse di questa molecola di DNA viene sottoposto a torsione in modo tale che il suo numero di contorcimento non sia più zero. (Non c'è alcun modo intuitivo per stimare il numero di contorcimento di una curva. Un nastro che si contorca nel senso comune di questa parola potrebbe risultare con un numero di contorcimento totale nullo.) Quando il numero di contorcimento della molecola viene modificato, il numero di legame rimane identico (può modificarsi solo se una delle due impalcature di sostegno della doppia elica si infrange) e l'avvolgimento è costretto a modificarsi. È questo rapporto che sta alla base del fenomeno dell'avvolgimento a superelica.

Il numero di legame, l'avvolgimento totale e il numero di contorcimento di un nastro non dipendono dalla localizzazione o dall'orientamento nello spazio. Sono anche indipendenti dalla scala, ma se un asse dello spazio viene invertito (come nel caso in cui il nastro viene riflesso in uno specchio) o addirittura tre assi sono invertiti (come nel caso in cui il nastro stesso sia invertito in corrispondenza di un punto), il segno di tutte e tre le quantità si modifica. D'altra parte, se due assi qualsiasi sono invertiti, come accade se si guarda in un comune microscopio (non dissezionatore), i segni rimangono tali e quali. Vi sono due altri casi di interesse pratico. Effettuando una stampa a contatto di una proiezione del nastro, i segni di *Lg*, *At* e *Co* si modificano. D'altra parte, se il negativo di una fotografia di questa proiezione viene posto correttamente in un ingranditore, la stampa che si otterrà mostrerà un nastro con i segni di tutte e tre le quantità immutati. Per contro, se il negativo nell'ingranditore viene capovolto, i segni di *Lg*, *At* e *Co* si invertiranno sulla stampa.

In breve, qualunque operazione che trasforma un avvitamento destrorso in uno sinistrorso senza introdurre altre distorsioni modificherà il segno del numero di legame, l'avvolgimento totale e il numero di contorcimento. Uno di noi (White) e Thomas F. Banchoff della Brown University hanno dimostrato che esiste un'altra operazione matematica speciale che modifica il segno di queste quantità, ma lascia la loro grandezza inalterata: l'inversione di un nastro attraverso una sfera. Questo risultato spiega perché il numero di contorcimento di un nastro il cui asse sia situato sulla superficie di una sfera è zero. Con quest'operazione, la curva chiusa descritta dall'asse si trasforma in se stessa.

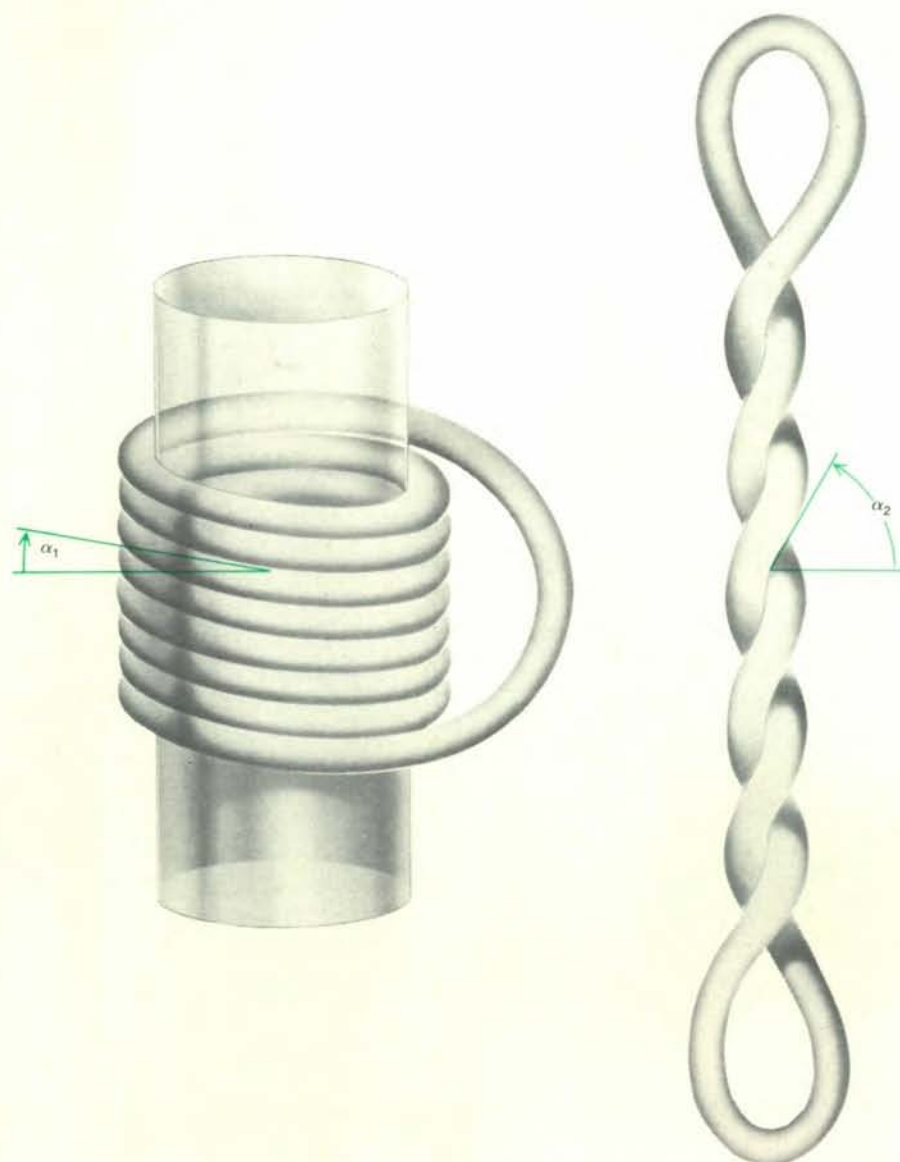
Benché l'equazione $Co = Lg - At$ dimostri che il legame e l'avvolgimento sono matematicamente distinti, la differenza fisica tra queste quantità può non

La relazione tra contorcimento e attorcigliamento si può dimostrare con un filo del telefono a spirale. Quando tale filo si trova nello stato di riposo *(in alto)*, si può avere uno scarso attorcigliamento e, invece, un sostanziale contorcimento: l'asse del filo traccia un'elica nelle tre dimensioni dello spazio. Quando il filo viene tirato *(al centro)* in modo che il suo asse risulti pressoché rettilineo *(in basso)*, il contorcimento diminuisce, mentre al contrario l'attorcigliamento diventa notevole.

Le principali configurazioni della doppia elica includono una forma lineare *(a sinistra)*, una forma circolare che presenta delle interruzioni e pertanto è decontratta *(al centro)* e una forma circolare chiusa (integra) e avvolta a superelica *(a destra)*. (Le tre molecole sono disegnate in scale diverse.) La conformazione della molecola circolare chiusa a superelica con avvolgimenti che si intersecano può essere spiegata in termini di rapporto tra legame, contorcimento e attorcigliamento. Dato che tale molecola risulta meno attorcigliata, essa ha un deficit nel numero di legame in confronto a una molecola che sia decontratta e che abbia la stessa dimensione e compensa questo fatto contorcendosi, avvolgendosi e piegandosi e soddisfacendo l'equazione $Co = Lg - At$.

essere ancora evidente. Può essere utile allora considerare ciò che accade quando un nastro matematico si avvolge attorno a un cilindro in modo tale che la sua superficie sia sempre piatta contro di esso *(si veda la figura a pagina 92)*. (Si noti che, solo per il fatto che un nastro matematico ha uno spessore infinitesimale, si riesce a farlo stare piatto in questo modo contro la superficie di un cilindro.) Chiameremo α l'angolo di inclinazione dell'elica descritto da questo nastro. In altre parole, α è l'angolo in corrispondenza del quale ogni giro dell'elica si inclina rispetto a un piano orizzontale, per cui quando è piccolo l'elica è schiacciata, quando è grande l'elica è invece molto ripida e allungata.

Si supponga ora che il nastro sia avvolto attorno al cilindro N volte prima che le sue estremità si congiungano nel modo più semplice. Allora, se si ignorano gli effetti finali, si può dimostrare che il numero di legame del nastro Lg è uguale a N, mentre l'avvolgimento totale At è uguale a N sen α. Pertanto, quando l'elica viene stirata in modo che l'angolo di inclinazione α aumenti, il numero di giri, e così anche il numero di legame, rimangono gli stessi; l'attorcigliamento passa, invece, da un valore basso a uno alto, dimostrando chiaramente la differenza tra legame e avvolgimento.

Dato poi che il numero di contorcimento di un nastro viene definito come la differenza tra numero di legame e avvolgimento totale, il suo valore per questo nastro è $N - N$ sen α, o $N(1 - \text{sen}\,\alpha)$. Come indica questa formula, quando α è piccolo e l'avvolgimento scarso, il contorcimento è notevole, mentre quando α è grande e l'avvolgimento sostanziale, il contorcimento è minimo. Si può osservare facilmente questa relazione con un filo del telefono a spirale: quando questo filo non è sottoposto a sollecitazioni, assume una forma con contorcimento elevato, ma con piccolo avvolgimento; quando, invece, le sue estremità vengono tirate, si ottiene una forma con notevole avvolgimento e, invece, basso contorcimento.

Si consideri in che modo questi risultati si possono applicare al DNA reale, ad esempio al DNA del virus del polioma. Si ricordi che questo DNA può essere risolto mediante sedimentazione in tre componenti: I e II, che sono circolari, e III, che è lineare. Si può determinare sperimentalmente che il numero di legame medio per una popolazione di molecole circolari decontratte di questo tipo di DNA sia circa +500. (Perfino in un preparato estremamente purificato, non tutte le molecole sono decontratte o avvolte a superelica nella stessa misura e così il numero di legame medio di una popolazione non è, necessariamente, un numero intero.) D'altra parte, per una popolazione di molecole circolari chiuse (le molecole a superelica che fanno parte del componente I), il numero medio di legame è circa +475. Proprio come aveva previsto Vinograd, le molecole circolari chiuse del DNA del virus del polioma sono meno attorcigliate, avendo nel loro numero di avvolgimento un deficit di 25. Questo risultato fa venire in mente un modo per definire l'avvolgimento a superelica. Esso è uguale a ΔLg, cioè alla differenza tra numero di legame di una molecola allo stato circolare chiuso, naturale, e il numero di legame della stessa molecola nello stato circolare chiuso decontratto (in cui l'energia di deformazione è ai valori minimi e il numero di contorcimento è zero). Per esempio, per i DNA di virus del polioma e di virus della scimmia SV 40, la ΔLg è circa -25. (Si ipotizza che il numero di legame venga misurato in soluzione sia per il DNA decontratto sia per il DNA avvolto a superelica, in condizioni standard di temperatura, di salinità e così via.)

Si può oggi capire perché un deficit del numero di legame di una molecola di DNA provochi l'avvolgimento a superelica di quest'ultima. Una molecola lineare di DNA in soluzione assume, normalmente, una forma nota come configurazione B, in cui le basi nucleotidiche sono all'incirca perpendicolari all'asse dell'elica, con unità di 34×10^{-9} centimetri tra loro e circa 10 coppie di basi per ogni giro di doppia elica. È questa una configurazione con energia minima e, se la molecola si ripiega o si attorciglia, tale energia aumenta. Se però una molecola lunga forma semplicemente un cerchio, il diametro di tale cerchio è grande in confronto allo spessore della doppia elica e, pertanto, la curvatura della molecola è piccola e la sua energia aumen-

Un tubo di gomma, che è stato avvolto a elica attorno a un cilindro e le cui estremità sono state congiunte in modo tale che la tensione dovuta all'avvolgimento si annulla *(a sinistra)*, passa - quando si toglie il cilindro - a una configurazione a elica, con avvolgimenti che si intersecano, e attorcigliata in direzione opposta *(a destra)*. Se, nella prima configurazione, ci sono N giri d'elica destrorsi e se l'angolo di inclinazione α_1, in corrispondenza del quale ogni giro d'elica si piega rispetto al piano orizzontale, è piccolo, allora, nella seconda configurazione vi saranno approssimativamente $N/2$ giri verso l'alto e $N/2$ giri verso il basso e il nuovo angolo di inclinazione α_2 sarà grande. (È solo parzialmente vero.) Esaminando i cambiamenti nei valori del numero di legame, dell'avvolgimento totale e del numero di contorcimento che accompagnano questa trasformazione, si vede perché la maggior parte dei DNA naturali, aventi un deficit nel numero di legame, si attorciglia in una superelica con avvolgimenti che si intersecano e che, come la doppia elica, è destrorsa.

ta solo lievemente. Come risultato, le molecole circolari «rotte» come quelle del componente II del DNA del virus del polioma, difficilmente si discostano dalla configurazione *B*.

La situazione è del tutto diversa per una molecola circolare chiusa con un deficit nel numero di legame. Perché il valore di *Lg* sia minore di quello di una molecola decontratta (per esempio 475 contro 500) si dovrebbe «svolgere» la doppia elica, una trasformazione che aumenterebbe in modo sostanziale l'energia di deformazione della molecola. Nella superelica, tuttavia, la molecola circolare chiusa minimizza il distacco dalla configurazione *B*.

Per essere più precisi, come ha rivelato l'analisi del modello a nastro, il DNA con un avvolgimento minore può ridurre la propria energia di deformazione, assumendo una forma contorta. Dato che contorcimento e avvolgimento sono interconvertibili, è evidente che, modificando l'entità del contorcimento, si può ridurre al minimo l'avvolgimento di una molecola, riducendo così al minimo anche la componente della sua energia di deformazione, dovuta all'avvolgimento. D'altra parte, il contorcimento introduce sempre una certa curvatura e aumenta così il contributo, dato dal ripiegamento, al contenuto energetico della molecola. Pertanto, la configurazione a superelica che la molecola di DNA meno compattamente avvolta assume rende minimo tale avvolgimento, e introduce la minima quantità possibile di ripiegamento.

Così, quando l'asse del DNA viene fatto contorcere, la doppia elica risponde attorcigliandosi e/o piegandosi. Le forme che questo avvolgimento a superelica può assumere variano da una superelica solenoidale sinistrorsa a una superelica destrorsa, con attorcigliamenti che si intersecano. In natura, questa seconda configurazione è generalmente preferita. Il «negoziato» tra contorcimento e avvolgimento, che porta preferenzialmente alla forma di avvolgimento che si interseca, può essere reso più evidente mediante un esempio meccanico. Si consideri un pezzo di tubo di gomma flessibile e spesso avvolto a elica sinistrorsa attorno alla superficie di un cilindro. Si ammetta che, mentre si avvolge, esso abbia anche libertà di attorcigliarsi attorno al proprio asse, per cui la sua energia di avvolgimento risulta zero. Le sue estremità vengono quindi avvicinate (in modo tale da ridurre qualsiasi tensione dovuta all'avvolgimento) e congiunte in maniera definitiva. Se, a questo punto, si toglie al tubo ogni costrizione, esso passerà alla conformazione a superelica destrorsa con avvolgimenti che si intersecano *(si veda la figura a pag. 94)*. Perché?

Senza intraprendere un'analisi particolareggiata della meccanica di questa situazione, si può acquisire una comprensione generale di come una superelica sinistrorsa, che ha uno scarso avvolgimento mentre è estremamente contorta, si trasforma in una superelica destrorsa con avvolgimenti che si intersecano, considerando il numero di legame, l'avvolgimento totale e il numero di contorcimento delle due configurazioni. (Per rendere chiaro in questo contesto il significato di numero di legame, si immagini una coppia di linee parallele, disegnate sui lati opposti del tubo di gomma non avvolto a elica.) Si ricordi che, nella prima configurazione, l'avvolgimento complessivo o totale era stato deliberatamente posto uguale a zero e così (ignorando gli effetti terminali) il numero di legame del tubo doveva essere uguale al numero di contorcimento. Nell'esempio del nastro avvolto a elica attorno a un cilindro, abbiamo osservato che il numero di contorcimento per tale configurazione è dato dalla formula $Co = -N_1(1 - \text{sen}\,\alpha_1)$, in cui N_1 è il numero di giri d'elica nel tubo e α_1 è l'angolo di inclinazione. (Il segno negativo deriva dal fatto che l'elica è sinistrorsa.) Pertanto, per la prima configurazione, il numero di legame è uguale a $-N_1(1 - \text{sen}\,\alpha_1)$. (Dato che si trascurano gli effetti finali, questo numero generalmente non sarà mai intero.)

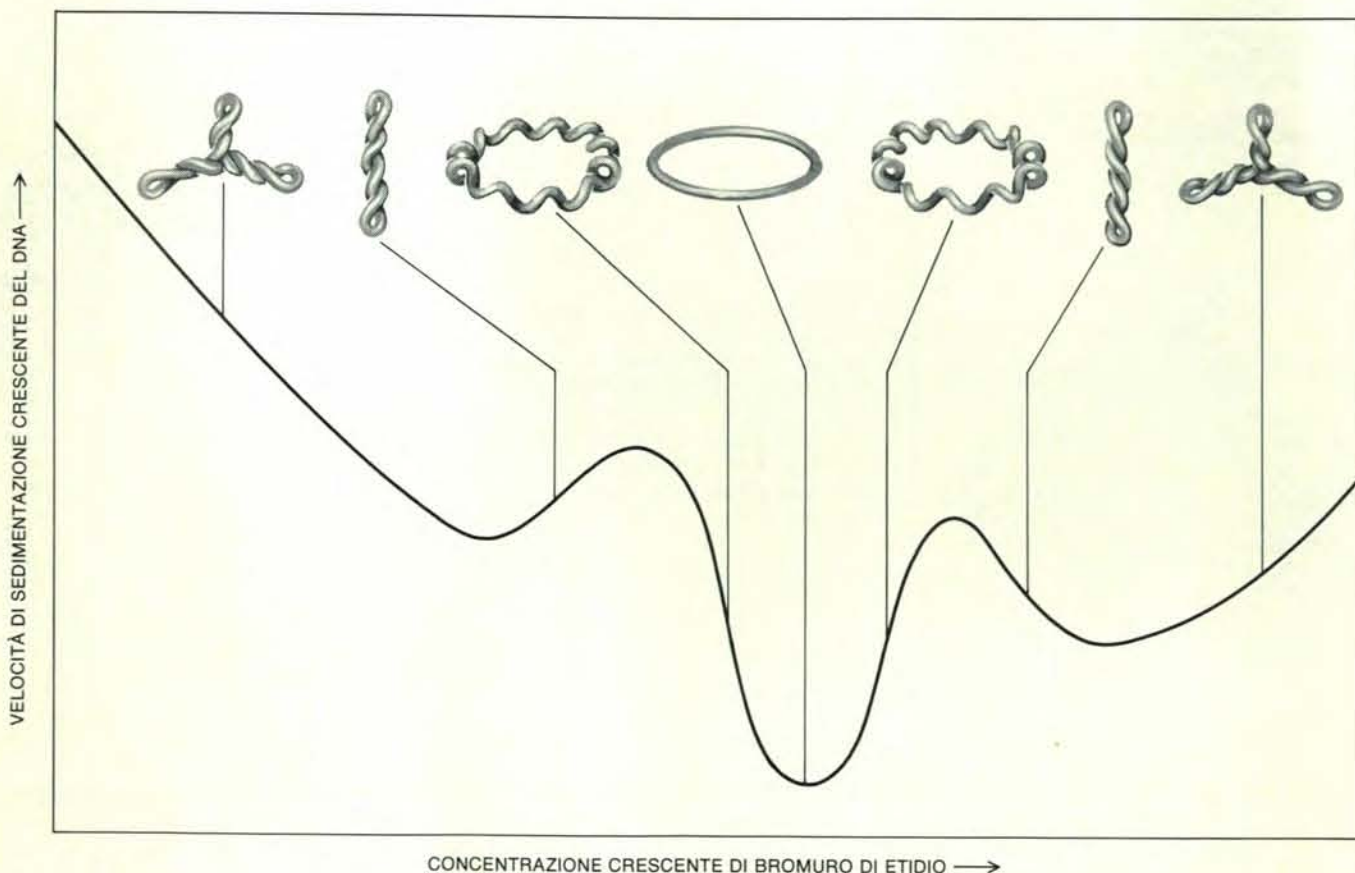
Rivolgendosi ora alla seconda configurazione del tubo di gomma, la superelica destrorsa con avvolgimenti che si intersecano, chiameremo il numero totale di giri destrorsi (in su e in giù) dell'elica N_2 e il suo angolo di inclinazione α_2. Pertanto (trascurando anche qui gli effetti finali) si può mostrare che, per una simile superelica, il numero di contorcimento *Co* è pari a $-N_2\,\text{sen}\,\alpha_2$. Si noti che la superelica sinistrorsa è stata trasformata nella superelica con avvolgimenti che si intersecano senza rompere il tubo di gomma o separare le estremità. Pertanto il numero di legame della seconda configurazione deve essere lo stesso della prima, cioè *Lg* deve essere uguale a $-N_1(1 - \text{sen}\,\alpha_1)$. Sostituendo questi valori di *Co* e di *Lg* nell'equazione $Co = Lg - At$, si ottiene $-N_2\,\text{sen}\,\alpha_2 = -N_1(1 - \text{sen}\,\alpha_1) - At$, oppure $At = N_1(1 - \text{sen}\,\alpha_1) + N_2\,\text{sen}\,\alpha_2$.

L'angolo di inclinazione della prima configurazione del tubo di gomma (α_1) era piccolo, mentre quello della seconda (α_2) era piuttosto grande. Pertanto, si possono semplificare ulteriormente le cose sostituendo nell'equazione le seguenti approssimazioni: $\text{sen}\,\alpha_1 = 0$ e $\text{sen}\,\alpha_2 = 1$, il che dà $At = N_1 + N_2$. Questa semplice formula spiega perché il tubo di gomma, lasciato libero, è passato a una configurazione con avvolgimenti che si intersecano e con direzione dell'attorcigliamento opposta a quella precedente. A queste approssimazioni, l'avvolgimento del tubo *At* è minimo quando i giri che esso compie nella seconda configurazione sono di numero uguale e avvengono in direzione opposta rispetto a quelli della prima configurazione. Inoltre, la curvatura della seconda configurazione è chiaramente inferiore a quella della prima. Pertanto, quando l'attorcigliamento è lieve, la superelica con avvolgimenti che si intersecano avrà un'energia di deformazione molto più bassa di quella della superelica semplice e il tubo di gomma assumerà, naturalmente, tale forma. (Per poter applicare questa argomentazione al DNA, dovremmo disegnare la doppia elica sul tubo di gomma, decontratto e non arrotolato, ma ciò non influirebbe sul calcolo dell'energia.)

Per poter effettuare un'analisi particolareggiata dell'avvolgimento a superelica del DNA reale, occorre avere una conoscenza piuttosto precisa di tutte le sue costanti elastiche ed è necessario considerare non solo gli effetti finali, ma anche fenomeni come la repulsione delle cariche e il moto termico. Tuttavia, le precedenti argomentazioni mostrano chiaramente perché una molecola di DNA circolare chiusa, con un deficit nel suo numero di legame (o, come è talvolta il caso, un eccesso) si avvolgerà in una superelica. In natura, la maggior parte del DNA circolare chiuso forma delle superieliche negative, cioè il suo avvolgimento a elica deriva da un deficit nel numero di legame. L'analisi del modello rappresentato dal tubo di gomma spiega perché ci si può aspettare che tali molecole assumano la configurazione di un'elica con avvolgimenti che si intersecano e il cui decorso - cosa piuttosto sorprendente - è lo stesso della doppia elica, cioè destrorso.

Le impalcature di sostegno zucchero-fosfato di una doppia elica hanno sufficiente flessibilità per cui, se in una molecola a superelica una di esse è rotta, l'altra può ruotare attorno a essa per porre la molecola in una configurazione decontratta. La molecola non intaccata non può perdere la sua configurazione a superelica in questo modo senza che un'impalcatura di sostegno integra passi attraverso l'altra, il che è fisicamente impossibile. Tuttavia, data una molecola di DNA naturalmente avvolta a superelica, con un deficit nel suo numero di legame, è possibile che il suo contorcimento si riduca senza che una delle due impalcature sia intaccata? Se le due impalcature devono rimanere intatte, allora il numero di legame non deve cambiare e, dato che il numero di contorcimento in questo caso è negativo, l'unico modo perché si abbia una riduzione nel contorcimento della molecola è che si abbia una riduzione dell'avvolgimento totale.

Se, a una soluzione di DNA, si aggiunge del bromuro di etidio, un colorante con una struttura molecolare planare, si ottiene questo scopo. Leonard S. Lerman della State University di New York a Albany è stato il primo a dimostrare che molecole di questo tipo si intercalano tra le coppie di basi della doppia elica. Nel DNA a superelica, questa inserzione provoca uno srotolamento locale della doppia elica e, in realtà, una concentrazione sufficientemente elevata di bromuro di etidio porterà la molecola alla decontrazione completa.

Per esprimersi in termini più precisi, questa concentrazione critica ridurrà l'avvolgimento totale della molecola. Per esempio, in presenza del colorante, un frammento lineare del DNA del virus del polioma avrebbe un avvolgimento totale di 475 e non di 500. Pertanto una molecola circolare chiusa di questo DNA, con un numero di legame 475, in presenza di questa quantità critica di colorante, avrebbe un avvolgimento totale uguale al suo numero di legame e, pertanto, il suo numero di contorcimento sarebbe zero. L'asse della molecola cesserebbe di contorcersi. Inoltre, se la concentrazione del bromuro di etidio venisse aumentata oltre il punto critico, l'avvolgimento totale verrebbe ulteriormente ridotto provocando così nella molecola circolare chiusa un

Quando il colorante bromuro di etidio, che ha una struttura molecolare planare, viene aggiunto a una soluzione di DNA a superelica, le superieliche si decontraggono come appare in questo grafico, che mostra in che modo la velocità di sedimentazione del DNA si modifica in seguito all'aggiunta di quella sostanza. La velocità di sedimentazione è la velocità con la quale le molecole di DNA si spostano attraverso il solvente quando la soluzione in cui si trovano è sottoposta a un campo gravitazionale intenso, generato da un'ultracentrifuga. Le molecole più strettamente avvolte a superelica sono più compatte e, pertanto, hanno una velocità di sedimentazione superiore. Aumentando la concentrazione del bromuro di etidio si riesce a diminuire gradatamente il contorcimento delle molecole di DNA, non riducendo il deficit del loro numero di legame (ciò si può effettuare solo creando delle intaccature nelle loro catene polinucleotidiche), ma riducendo il loro avvolgimento. Le molecole del bromuro di etidio vengono intercalate tra le coppie di basi del DNA, creando uno srotolamento locale della doppia elica. Il grafico mostra che, con una concentrazione abbastanza elevata del colorante, le molecole di DNA si decontraggono del tutto; aggiungendone ancora, il DNA comincia ad avvolgersi a superelica in direzione opposta. Questi cambiamenti di struttura, sono confermati dalla microscopia elettronica.

contorcimento che la porterebbe ad assumere la configurazione di una superelica con orientamento opposto.

Questi effetti si possono osservare direttamente studiando la velocità di sedimentazione delle molecole di DNA, come è stato mostrato la prima volta da Lionel V. Crawford dell'Imperial Cancer Research Institute in Gran Bretagna e da Michael J. Waring della Università di Cambridge; inoltre, indipendentemente, da Vinograd e da uno di noi (Bauer). L'aggiunta di bromuro di etidio a una soluzione di DNA circolare chiuso di virus del polioma fa sì che tali molecole si contorcano meno, per cui diventano meno compatte e sedimentano più lentamente. Aumentando la quantità di bromuro di etidio aggiunta, la velocità di sedimentazione raggiunge un minimo; quindi, dato che concentrazioni ancora maggiori di colorante fanno avvolgere le molecole in una superelica che ha un orientamento opposto e che le fa diventare di nuovo più compatte, la velocità di sedimentazione comincia ancora a salire.

Per creare molecole di DNA con diversi gradi di attorcigliamento, sarebbe utile poter provocare una intaccatura in una delle due impalcature zucchero-fosfato di una doppia elica di DNA, far decontrarre la molecola di alcuni giri e quindi riparare l'intaccatura. Cosa sorprendente, sono stati identificati gli enzimi in grado di fare proprio questo. Il primo, un enzima noto come proteina ω, è stato scoperto, nel 1971, in *Escherichia coli* da James C. Wang della Harvard University. Enzimi analoghi sono stati scoperti in numerose altre fonti, tra cui diversi batteri, il virus vaccinico e il nucleo e i mitocondri di cellule animali. Questi enzimi, che riparano le intaccature e sono chiamati topoisomerasi, in genere non richiedono fonti energetiche per funzionare. Agiscono sempre riducendo l'avvolgimento a superelica di una molecola di DNA e pertanto ne abbassano l'energia.

Molecole di DNA, con un avvolgimento a super-superelica, cioè con un maggior numero di giri di superelica di quanti siano normalmente presenti possono essere ottenute impiegando del bromuro di etidio *(punti in colore)* assieme a un enzima che serve a risanare le rotture, enzima che agisce riducendo l'avvolgimento a superelica di una molecola di DNA e provocando innanzitutto una temporanea frattura in una delle sue impalcature di sostegno zucchero-fosfato, e in seguito facendo decontrarre la molecola e quindi richiudendo la frattura. Pertanto, se viene aggiunta alla soluzione di DNA a superelica una sufficiente quantità di bromuro di etidio *(a)* perché il DNA possa decontrarsi completamente *(b)* e quindi possa riavvolgersi a superelica in direzione opposta *(c)*, l'aggiunta dell'enzima che salda le fratture (topoisomerasi) riporterà il DNA allo stato decontratto *(d)*. Tuttavia, una quantità sostanziale di colorante si legherà al DNA per cui questo si troverà in una situazione di minor avvolgimento rispetto a quella usuale. Pertanto, se l'enzima prima e il colorante poi vengono eliminati, la molecola di DNA tornerà ad avvolgersi di nuovo a superelica in modo ancora più stretto dell'usuale e orientata nella direzione originaria *(e)*.

Le topoisomerasi si sono dimostrate strumenti di incalcolabile valore per studiare la chimica-fisica del DNA. Tanto per cominciare, l'aggiunta di una topoisomerasi a una soluzione di DNA a superelica riduce gradatamente l'avvolgimento fino a quando tutte le molecole di DNA si trovano nello stato decontratto o quasi. Tuttavia, applicazioni ben più ingegnose sono state studiate per questi enzimi. Per esempio, si supponga che una concentrazione elevata di bromuro di etidio venga aggiunta a DNA circolare chiuso, provocandone l'avvolgimento a superelica nella direzione opposta a quella abituale. Se viene aggiunta anche una topoisomerasi, le molecole in soluzione si decontraggono, ma a causa della grande quantità di bromuro di etidio intercalato saranno notevolmente meno attorcigliate che nel loro stato naturale. A questo punto, se viene tolta prima la topoisomerasi e quindi viene eliminato il colorante, le molecole compenseranno l'ulteriore riduzione del numero di legame, avvolgendosi ancora più compattamente a superelica nella direzione abituale. In questo modo si possono ottenere delle molecole a super-superelica, cioè molecole che hanno un avvolgimento a superelica molto superiore al normale. Questa forma di immagazzinamento e di trasferimento di energia potrebbe essere il perno della funzione che l'avvolgimento a superelica avrebbe nella cellula. Si potrebbe perfino concepire un dispositivo chimico che, passando attraverso un ciclo analogo di operazioni, trasformasse in lavoro l'energia chimica.

È anche interessante notare che Martin Gallart e collaboratori dei National Institutes of Health hanno trovato nel batterio *Escherichia coli* un enzima che, data una fonte di energia come l'adenosintrifosfato (ATP), ridurrà il numero di legame del DNA circolare decontratto, aumentando così il suo numero di contorcimento e facendolo attorcigliare a superelica. L'attività di un enzima di questo tipo, detto girasi, è essenzialmente opposta a quella delle topoisomerasi. Finora le girasi sono state trovate in una varietà di microrganismi, ma mai negli organismi superiori.

Nicholas R. Cozzarelli e collaboratori della Università di Chicago hanno dimostrato di recente che la girasi agisce apparentemente avvicinando due segmenti della molecola di DNA. Quindi taglia

La tecnica sperimentale per misurare l'avvolgimento a superelica comporta lo spostamento per elettroforesi su un gel di agarosio, in questo caso dall'alto in basso, di una popolazione di molecole di DNA circolare chiuso. Dato che le molecole maggiormente avvolte a superelica risultano più compatte, esse si muovono più velocemente attraverso il gel. Di conseguenza, si forma una successione di bande nettamente distinte, in modo tale che le molecole in ogni banda abbiano tutte lo stesso numero di legame e i numeri associati con le bande adiacenti differiscano di un valore di 1. (Il DNA rappresentato qui è stato colorato con un colorante fluorescente e, pertanto, le bande contenenti più molecole sono più luminose.) Se la popolazione comprende molecole con vari gradi di avvolgimento a superelica, come qui, il conteggio delle bande che si formano fornisce una misura diretta del deficit nel numero di legame che determina il grado di avvolgimento. La banda più in alto nel gel contiene il DNA rotto.

ambedue le impalcature zucchero-fosfato di uno dei due segmenti e passa il segmento intatto attraverso la lacuna formatasi prima di ridurre le impalcature tagliate. È facile mostrare che un simile processo altererebbe il numero di legame del DNA con incrementi di due unità, invece che una, e questo è, esattamente, quello che è stato osservato sperimentalmente.

Come si misura la formazione di una superelica? Un metodo sperimentale è basato su una scoperta compiuta nel 1974 da Walter Keller dell'Università di Heidelberg. Egli ha osservato che, sottoponendo a elettroforesi su un gel di agarosio molecole a superelica con numeri di legame anche lievemente diversi, queste molecole non si spostano tutte alla stessa velocità. Quelle più compatte si muovono più velocemente e, pertanto, si forma nel gel una successione di bande ben individualizzate. Inoltre, come è stato mostrato in seguito da Vinograd, se una di queste bande viene tagliata dal gel e le molecole che la compongono vengono nuovamente sottoposte a elettroforesi, ne risulta di nuovo una unica banda esattamente nella stessa posizione. E, come illustrato in epoca posteriore da Wang e da uno di noi (Bauer), se questa banda viene riscaldata quasi al punto di fusione (in modo che le basi nelle doppie eliche si disaccoppino) e quindi fatta raffreddare lentamente (in modo che le doppie eliche si ricongiungano), formerà ancora un'unica banda nella stessa posizione.

Pertanto, le molecole in bande adiacenti differiscono di una entità che sopravvive al riscaldamento e al raffreddamento e, essendo le bande isolate, è quantizzata. Essa non può essere che il numero di legame. La sola ragionevole interpretazione è che i numeri di legame delle molecole presenti sul gel, in bande adiacenti, differiscano di un valore di 1. Pertanto, quando una popolazione di molecole di DNA, che include molecole che si trovano in tutti gli stati, dal decontratto all'avvolto a superelica, viene sottoposta a elettroforesi su gel, la ΔLg, cioè la differenza nel numero di legame che determina l'avvolgimento, può essere valutata direttamente contando le bande. Questo valore, determinato sperimentalmente, può essere utilmente sfruttato. Per esempio, lo si può utilizzare per calibrare l'effetto di srotolamento del bromuro di etidio sulla doppia elica. Risulta che ogni molecola intercalata di questo composto produca uno srotolamento locale di circa 26 gradi; il bromuro di etidio può così essere utilizzato come uno standard sussidiario per misurare il valore di ΔLg per qualsiasi molecola a superelica. (Si dovrebbe anche notare che l'effetto sorprendente dell'enzima girasi sul numero di legame del DNA può essere chiaramente osservato in esperimenti su gel.)

Uno dei risultati più importanti emersi finora dallo studio del DNA a superelica è la prova diretta che le due impalcature zucchero-fosfato del DNA si avvolgono realmente l'una attorno all'altra a intervalli di circa 10 coppie di basi. A questo proposito gli esperimenti più eleganti sono quelli compiuti da Wang. Egli ha confrontato molecole simili di DNA circolare, costruite con le tecniche dell'ingegneria genetica e ciascuna con circa 5000 coppie di basi. La loro lunghezza non era nota, ma si conosceva, invece, con esattezza la differenza di lunghezza tra due qualsiasi di queste molecole. Alcune differivano soltanto per una sola coppia di basi, altre per anche 400 coppie.

Le differenze nell'entità dell'avvolgimento a superelica, osservate quando queste molecole di diversa dimensione venivano fatte decontrarre da una topoisomerasi e quindi sottoposte a elettroforesi su gel fianco a fianco, hanno permesso a Wang di dedurre che il numero di coppie di basi per ogni giro della doppia elica doveva essere 10,4 più o meno 0,1. L'esperimento da lui compiuto è una bella dimostrazione della precisione e della sofisticazione delle moderne tecniche di biologia molecolare. Inoltre, sono stati proposti di recente parecchi modelli di DNA in cui le due catene polinucleotidiche non si avvolgono l'una attorno all'altra per formare una doppia elica, ma giacciono invece fianco a fianco per la maggior parte della loro lunghezza, avendo solo pochi giri d'elica. I risultati di Wang, con quelli di esperimenti affini, mostrano che questi nuovi modelli non devono essere giusti. Questa argomentazione topologica è molto drastica perché elimina tutti i modelli a catene affiancate, indipendentemente dai dettagli molecolari.

Come compare l'avvolgimento a superelica? Nel virus SV40 e nel virus del polioma, il DNA è avvolto a superelica perché, in generale, non è nudo nel nucleo della cellula. Durante la duplicazione, la doppia elica si avvolge sui nucleosomi (sferette proteiche che consistono di otto molecole istoniche con una o due molecole associate). Quando la molecola di DNA si stacca da questa struttura di sostegno si avvolge a superelica (in maniera molto simile a quella del tubo di gomma di cui si è parlato prima). Il fatto che il numero di legame sia sempre ridotto nei DNA naturali implica che, quando il DNA si avvolge attorno ai nucleosomi formando un'elica solenoidale, quest'elica è sinistrorsa. Incertezze nei dati che sono disponibili e argomentazioni sulla loro interpretazione rendono difficile il determinare il numero esatto di superieliche che vengono generate da ogni nucleosoma, ma sembra probabile che questo numero sia compreso tra 1 e 2. I nucleosomi sono associati con il DNA in tutti gli organismi superiori; in quelli inferiori l'origine della formazione di supereliche non è del tutto chiara.

Molto lavoro rimane ancora da compiere prima che tutte le implicazioni dell'avvolgimento a superelica del DNA possano essere comprese. In particolare, gli enzimi in rapporto con tale avvolgimento (topoisomerasi e girasi) e il ruolo di quest'ultimo nella duplicazione del DNA sono oggi oggetto di intense ricerche.

Le eliche

Le eliche marine e quelle aeree hanno un aspetto molto diverso, anche se sono progettate tutte secondo la tradizionale teoria della circolazione, sviluppata al fine di spiegare la portanza delle superfici aerodinamiche

di E. Eugene Larrabee

Nel 1845, agli albori della navigazione a vapore, la superiorità dell'elica nella propulsione navale fu dimostrata in una «gara di tiro alla fune» predisposta dall'Ammiragliato britannico fra la corvetta a vapore *Rattler*, dotata di tale propulsore, e *Alecto*, equipaggiata con ruote a pale, nella quale la prima rimorchiò la seconda di poppa a una velocità di quasi tre nodi. Le due navi, di caratteristiche molto simili, erano azionate da motori a vapore a basso numero di giri che sviluppavano una potenza di circa 200 cavalli-vapore (HP). La prova fu significativa in quanto dimostrò la preminenza della propulsione a elica rispetto al concetto della ruota a pale, di più intuitiva comprensione.

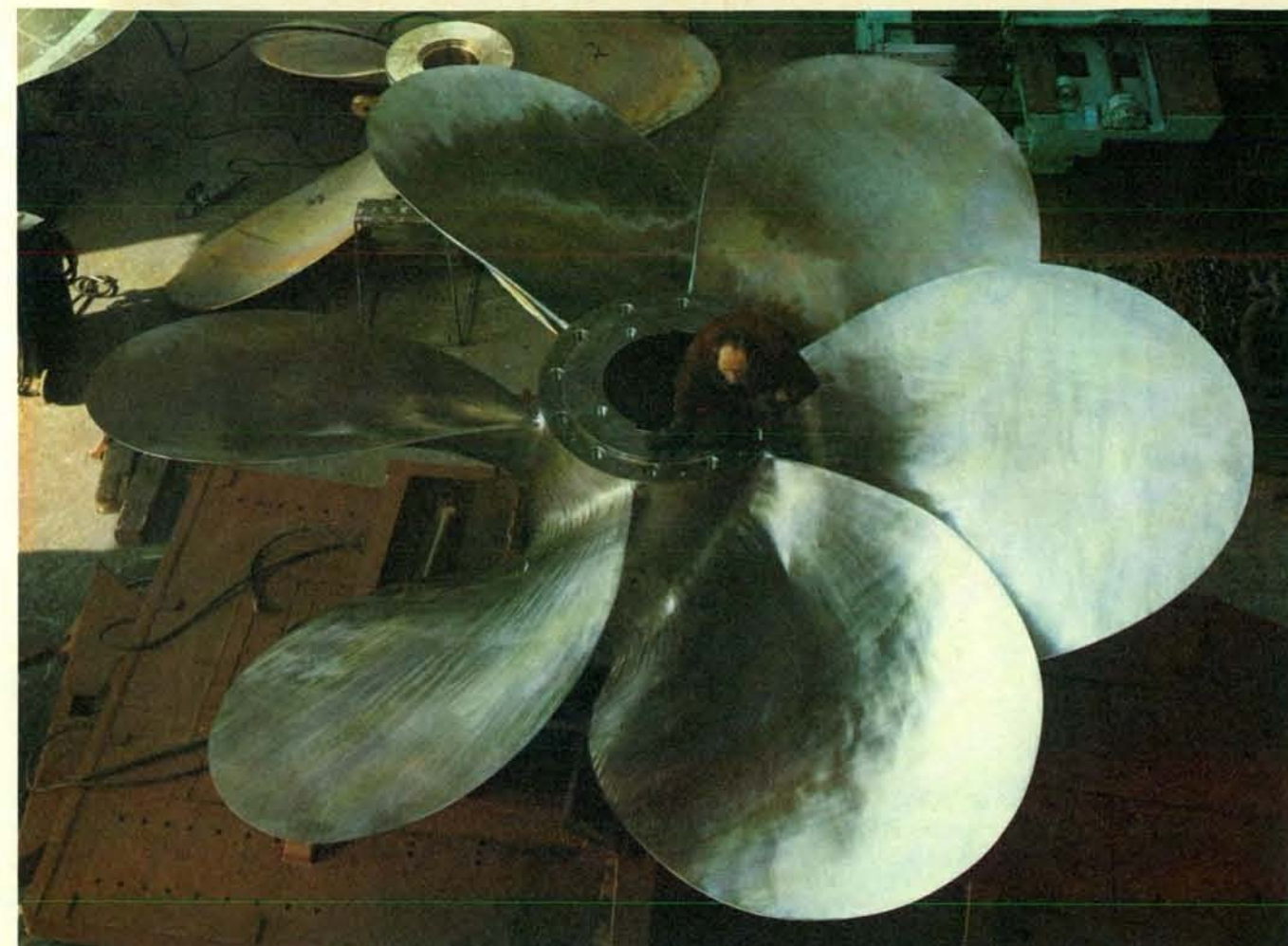
In campo navale l'elica è rimasta il mezzo di impiego più generale per la propulsione di imbarcazioni di qualsiasi tipo e dimensione. Le principali eccezioni sono rappresentate da alcuni natanti per impieghi speciali quali quelli a basso pescaggio o gli aliscafi. Quando poi si tratta di spingere un aeroplano con motore a pistoni non esistono alternative all'elica. Questo propulsore è usato anche sui velivoli a turboelica con motori a turbina. La maggior parte degli aerei commerciali di linea è oggi però azionata da motori a reazione (o a turbogetto) nei quali il dispositivo di propulsione assume un nome differente. In un motore a reazione il gas che esce ad alta velocità dalla camera di combustione passa attraverso una turbina motrice azionante un «ventilatore» ossia un compressore assiale disposto davanti al motore. Il compressore in realtà altro non è che un'elica dotata di moltissime pale realizzata in modo da funzionare all'interno della cassa del motore; esso fornisce l'aria per la combustione e nel contempo crea un flusso a bassa velocità e ad alto volume che segue una via di by-pass alla camera di combustione e si mescola poi con i gas di scarico caldi provenienti dalla turbina per imprimere la spinta al velivolo. Nel motore a turboelica la turbina aziona un'elica visibile all'esterno che fornisce la maggior parte della spinta. Come vedremo, il motore a turboelica potrebbe nuovamente diffondersi grazie al suo rendimento.

Chiunque abbia osservato attentamente un'elica marina e una tipica elica aerea si sarà reso conto che le pale della prima coprono una frazione molto maggiore dell'intero disco battuto durante la rotazione dalle pale di quanto non avvenga nel caso dell'elica per aerei. L'ingegnere aeronautico direbbe che l'elica marina ha una «solidità» maggiore. La differenza più evidente può vedersi nel confronto fra l'elica marina o «ruota» di una superpetroliera e l'elica del *Gossamer Albatross*, l'aeroplano ultraleggero azionato a pedali, con cui Bryan Allen ha trasvolato il Canale della Manica il 12 giugno 1979. Le eliche delle più grosse petroliere hanno cinque o sei pale, raggiungono diametri di nove metri, ruotano a 95 giri al minuto assorbendo una potenza di 45 000 HP, destinata a imprimere a navi di 500 000 o più tonnellate di stazza lorda una velocità di 16 nodi (8,23 metri al secondo). Le eliche gemelle della *Queen Elizabeth 2*, che raggiunge una velocità di 31 nodi, assorbono ciascuna 55 000 HP, sono notevolmente più piccole e hanno una solidità anche maggiore delle eliche delle più grosse petroliere.

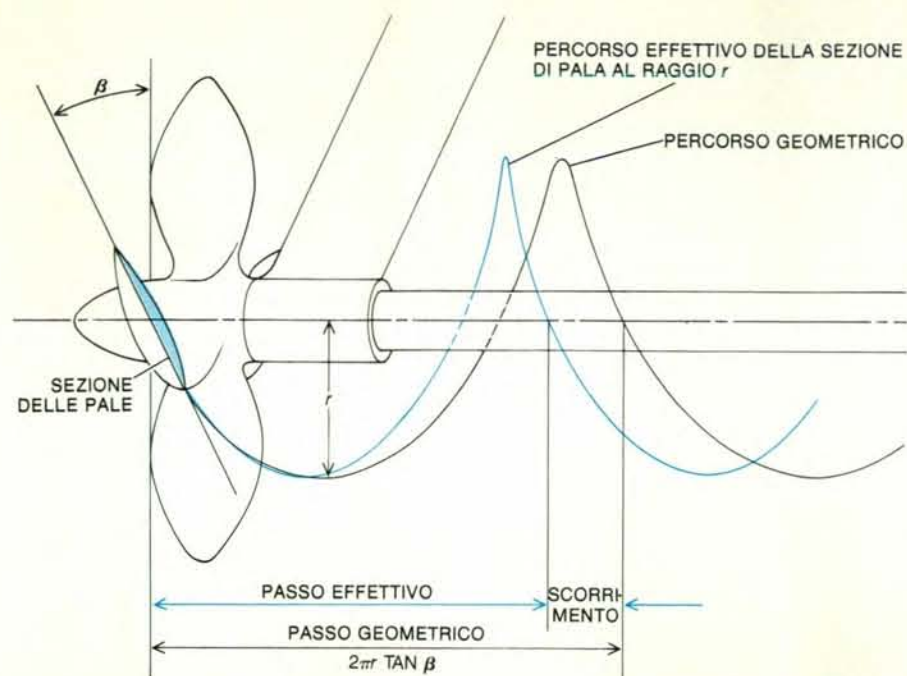
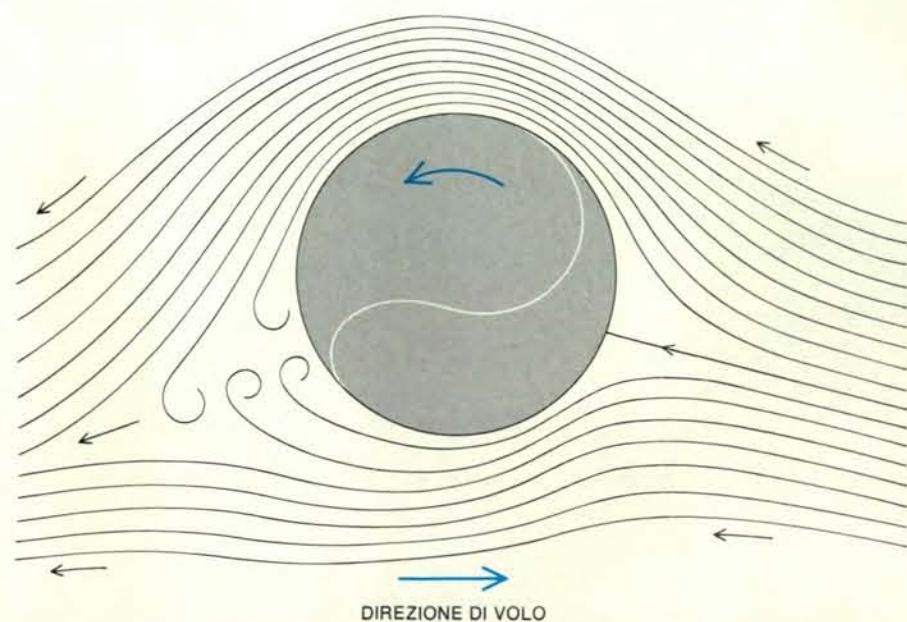
L'elica del *Gossamer Albatross* ha due sottili pale e un diametro pari al 70 per cento delle eliche della *Queen Elizabeth 2*. Quando quest'elica da 4,1 metri ruota a 130 giri al minuto assorbe 0,25 HP con cui spinge un aereo di 96 chilogrammi in assetto di volo alla velocità di 5,4 metri al secondo. Le eliche marine e aeree sono entrambe il risultato di analoghe analisi al calcolatore, che tengono conto della densità del fluido, dell'interazione dei campi di flusso, del carico idrodinamico o aerodinamico sulle pale e infine dei vincoli pratici imposti dal veicolo e dalle sue caratteristiche di funzionamento. Ma perché la geometria dei due tipi di elica è così diversa?

Il motivo non risiede nel fatto che l'aria è un fluido gassoso e l'acqua un fluido liquido. Per quanto possa sembrare strano, l'elica del *Gossamer Albatross* è molto più prossima alla forma ideale che dovrebbe avere un'elica di rendimento massimo adatta a lavorare in un fluido quasi incomprimibile a bassa viscosità (cioè un fluido come è l'acqua) di quanto non lo sia l'elica marina. La sua forma corrisponde alle considerazioni teoriche espresse nel 1919 da Albert Betz e Ludwig Prandtl del Kaiser Wilhelm Institut per le ricerche sulla dinamica dei fluidi di Göttingen e perfezionate nel 1929 da Sydney Goldstein, uno scienziato inglese in visita a tale istituto. Le loro idee presero le mosse dalla rivoluzione provocata nell'idrodinamica teorica, all'inizio del secolo, da Prandtl, con l'enunciazione della teoria dello strato limite, che permetteva di spiegare la resistenza all'avanzamento dei corpi aerodinamici e quando W. M. Kutta in Germania e Nikolai Joukowsky in Russia inventarono indipendentemente la teoria della circolazione per la portanza, che spiegava l'azione di sostentamento delle ali e la spinta impressa dalle pale delle eliche.

Fino ad allora l'idrodinamica teorica, che era stata sviluppata per studiare la resistenza all'avanzamento degli scafi, era stata soprattutto una disciplina accademica la quale si occupava dei diagrammi matematici di flusso che si potevano formare in varie combinazioni di campi di flusso. Le linee di flusso di alcuni dei campi matematici teorici potevano essere rese simili alla corrente intorno a uno scafo, ma Jean Le Rond d'Alembert aveva dimostrato nel 1742 che la resistenza teorica della forma matematica di scafo a una corrente del genere era esattamente zero. Inoltre, se la forma dello scafo veniva straorzata matematicamente (cioè fatta rotare di un certo angolo rispetto alla direzione del campo di flusso) anche la forza idrodinamica perpendicolare alla direzione di spostamento dello scafo risultava nulla. Poiché gli scafi reali non presentano resistenza nulla ai flussi effettivi, queste conclusioni divennero note come il paradosso di d'Alembert.

In assenza di teorie utilizzabili, nel diciannovesimo secolo le costruzioni navali divennero una scienza sperimentale e le

Una delle due eliche della *Queen Elizabeth 2* che insieme assorbono una potenza di 110 000 HP per muovere la nave di linea, dislocante 67 000 tonnellate di stazza lorda a una velocità massima di 31 nodi. L'elica ha un diametro di 5,8 metri e genera la spinta massima quando ruota a 174 giri al minuto. Il diametro dell'elica è limitato dal vincolo di non sporgere al di sotto della chiglia o al di sopra della linea di galleggiamento. Per raggiungere la spinta voluta con la necessaria limitazione di diametro, l'area complessiva delle pale è pari all'89,5 per cento dell'area del disco battuto. Nelle costruzioni navali il rapporto fra la superficie delle pale e quella del disco è denominato rapporto dell'area sviluppata; in aeronautica è detto solidità. L'elica della *Queen Elizabeth 2* è stata progettata e costruita dalla SMM Propeller Ltd. di Liverpool.

L'elica del *Gossamer Albatross*, l'aeroplano ultraleggero azionato a piedi con cui Bryan Allen ha trasvolato il Canale della Manica il 12 giugno 1979, presenta solo una vaga somiglianza con l'elica della *Queen Elizabeth 2* illustrata nella figura in alto. Tuttavia entrambe le eliche funzionano come «elemento di trazione» e sono fra i dispositivi di miglior rendimento del loro tipo mai progettati. L'elica aerea del *Gossamer Albatross* ha un diametro di 4,1 metri (l'80 per cento dell'elica della nave), ruota a 95 giri al minuto e assorbe 0,8 HP al decollo e 0,25 HP in volo di crociera. L'elica è stata progettata al calcolatore impiegando l'algoritmo ideato dall'autore al fine di presentare la minima energia dissipata nella scia vorticosa e le minime perdite per attrito nel volo in leggera salita. Il *Gossamer Albatross* è stato progettato dalla AeroVironment Incorporated di Pasadena in California. La fotografia è stata scattata al Johnson Space Flight Center di Houston nel Texas.

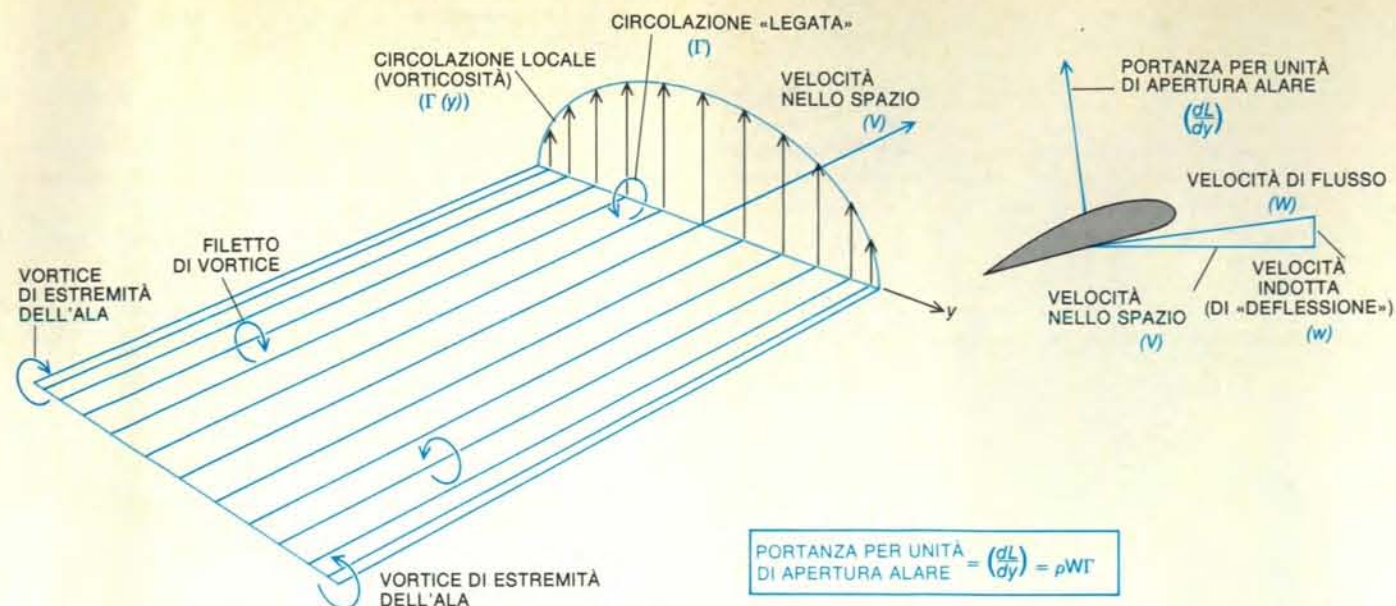
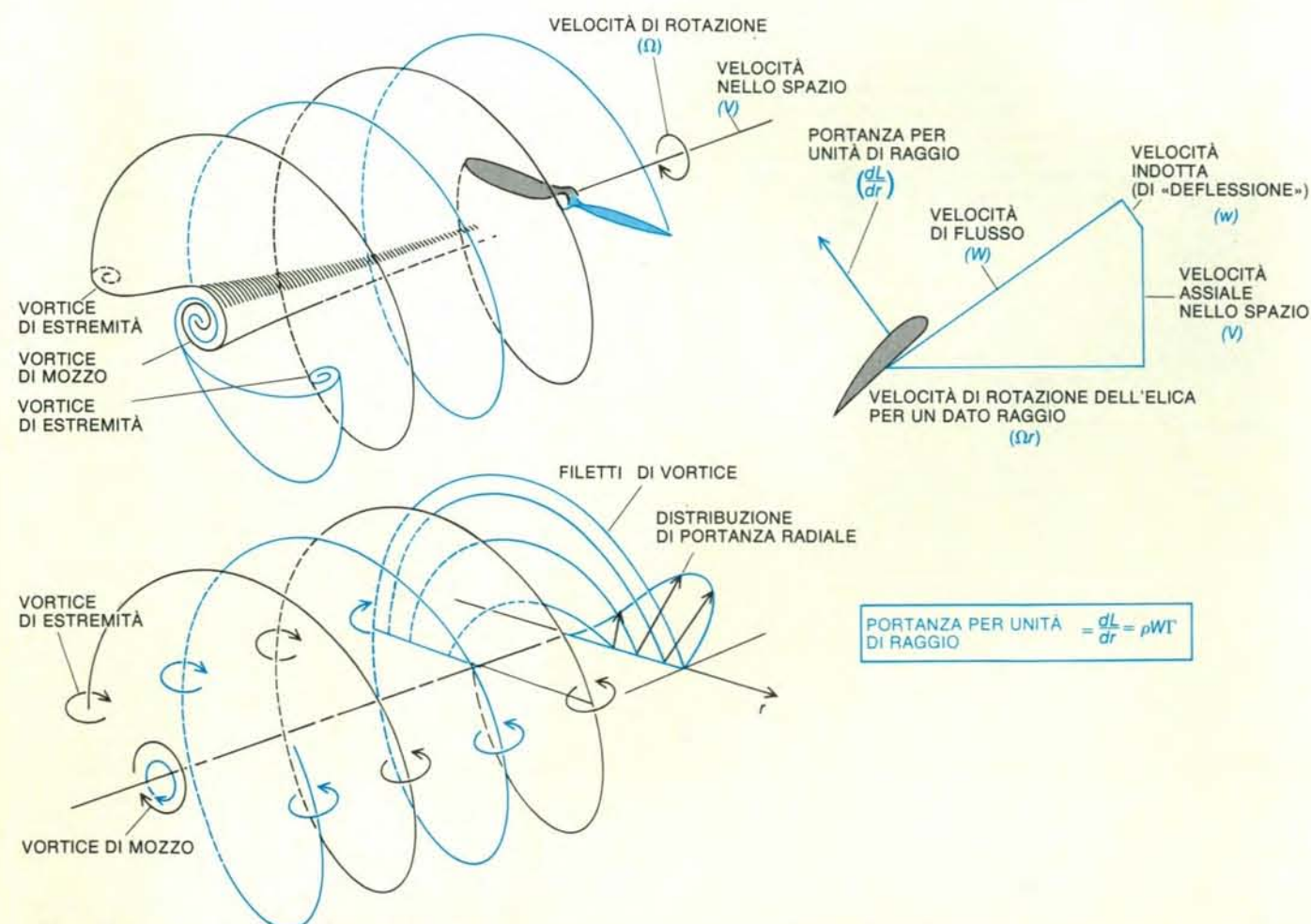
eliche marine furono sviluppate in modo intuitivo. L'elica venne assimilata a una vite che avanza ruotando in un foro filettato. Tuttavia al contrario del metallo l'acqua cede sotto la spinta dell'elica: pertanto il «passo effettivo», cioè la distanza coperta in un giro, è inferiore al «passo geometrico» dell'elica di una quantità chiamata scorrimento. Nel 1865 l'ingegnere scozzese William J. M. Rankine sviluppò una teoria dell'elica (denominata teoria del «disco attuatore»), più tardi perfezionata da William Froude, secondo la quale lo scorrimento si poteva spiegare come la conseguenza di un semplice aumento della quantità di moto del fluido nella scia dell'elica, un aumento proporzionale alla spinta di quest'ultima.

Il fenomeno dello scorrimento nelle eliche marine è stato individuato fin dalle prime applicazioni. La distanza coperta da una nave durante un giro dell'elica, il «passo effettivo», è inferiore al passo geometrico determinato dall'angolo delle pale dell'elica. William J. M. Rankine e William Froude hanno spiegato che lo scorrimento corrisponde all'aumento della quantità di moto assiale impartita dall'elica alla scia e, quando è eccessivo, dà luogo a un basso rendimento.

Una palla da tennis cui è stata impressa una rotazione contraria al moto di avanzamento imita quanto avviene su un profilo aerodinamico che distorce il campo di flusso in modo da creare una portanza aerodinamica. Forse la prima spiegazione del moto di una palla da tennis «tagliata» è attribuita a Lord Rayleigh. A causa della rotazione l'aria che fluisce sopra la palla viene accelerata, mentre quella che scorre sotto la palla risulta ritardata. Secondo il teorema del matematico svizzero Daniel Bernoulli, un enunciato della conservazione dell'energia, la pressione nella zona di flusso accelerato al di sopra della palla deve diminuire rispetto alla pressione nella zona di flusso ritardato sotto la palla. Il risultante squilibrio delle forze spinge la palla verso l'alto.

Secondo tale teoria, lo scorrimento è pari alla metà dell'aumento di velocità della scia (che dipende dalla spinta per unità di superficie di disco battuto e dalla velocità del natante) moltiplicato per il tempo impiegato dall'elica per compiere un giro. Lo scorrimento geometrico deve essere aumentato corrispondentemente quando si cerca di pervenire alla determinazione dello scorrimento effettivo (o desiderato), che è pari al rapporto fra la velocità della nave espressa in metri al minuto e la velocità dell'albero motore in giri al minuto. Sebbene quanto precede sia quasi vero in base all'esperienza, la teoria del disco attuatore non forniva alcuna spiegazione al fatto che le pale sviluppano soprattutto carichi idrodinamici di spinta e di torsione. Tuttavia essa rappresentava un precedente all'idea della velocità di «afflusso», che sarebbe stata identificata più tardi con la velocità «indotta» nella teoria delle ali, e collegava uno scorrimento eccessivo (assieme agli alti carichi) con un rendimento ridotto.

Una spiegazione dei carichi di spinta e di torsione attendeva lo sviluppo della teoria dei «sistemi di vortici» o della circolazione, una teoria della portanza suggerita da Kutta e Joukowsky fra il 1902 e il 1911. Secondo l'ipotesi di Joukowsky il flusso ideale di un liquido non viscoso intorno a un cilindro rotante può essere rappresentato in modo conforme ovvero trasformato nel corrispondente flusso intorno a una superficie aerodinamica avente un bordo di uscita sottile. La vorticosità o circolazione viene aggiunta al flusso intorno al cilindro introducendo un «limite» o «linea» ideale, cioè un vortice di intensità arbitraria coincidente con il centro del cilindro. L'intensità del vortice viene regolata con criteri matematici fin quando il flusso intorno al cilindro ristagna in un punto di singolarità che nella trasformazione coincide con il bordo di uscita della superficie aerodinamica.

La circolazione teorica del fluido non viscoso intorno a un cilindro di lunghezza infinita assomiglia superficialmente al flusso di un fluido reale intorno a una sfera rotante avente una superficie ruvida, come una palla da tennis o da golf, cui sia stata impressa una rotazione contraria al moto di avanzamento. La spinta si determina in conseguenza dell'accelerazione dell'aria che scorre sopra la parte superiore della sfera e della decelerazione di quella che scorre sotto la sfera. Secondo un ben noto teorema enunciato dal matematico svizzero Daniel Bernouilli nel diciottesimo secolo, il fluido nelle zone di flusso accelerato ha una pressione statica inferiore alla pressione delle zone di flusso decelerato, e quindi la sfera risulta spinta verso l'alto (una palla da tennis cui viene impressa una rotazione secondo il moto di avanzamento, riceve una corrispondente spinta dall'alto verso il basso). Sebbene i campi di flusso in un fluido non

La teoria dei vortici per la portanza è stata enunciata fra il 1902 e il 1911 in modo indipendente dal tedesco W. M. Kutta e dal russo Nikolai Joukowsky. Secondo la formulazione di Joukowsky, rappresentata in questa figura, un flusso di un fluido non viscoso intorno a un cilindro di lunghezza infinita può essere «trasformato» matematicamente nel flusso intorno a un profilo aerodinamico. Il flusso può essere simmetrico, come mostrato. Joukowsky aggiunse, al flusso del fluido non viscoso, un flusso vorticoso ideale coincidente con l'asse del cilindro. La intensità di questo vortice «di confine» o circolazione Γ, viene scelta in modo da dar luogo a un ristagno nel flusso in corrispondenza del punto singolare posteriore x sul cilindro. Quando la circolazione viene rappresentata sulla superficie aerodinamica, il campo di flusso è allineato con il bordo di uscita. In queste condizioni la circolazione crea una portanza sia sul cilindro sia sulla superficie aerodinamica. La portanza per unità di apertura alare è $\rho V \Gamma$, ove ρ è la densità del fluido (in questo caso aria) e V è la velocità della corrente remota, o non perturbata. Lo spessore della superficie aerodinamica è controllato dal valore assegnato alla dimensione e di sfalsamento. L'angolo di attacco della superficie aerodinamica, α, corrisponde al disallineamento dei due punti singolari della trasformazione, x e x', rispetto alla direzione della corrente remota.

Il concetto dello strato vorticoso di uscita è stato introdotto da Ludwig Prandtl per spiegare l'andamento del flusso intorno alle ali dove la portanza e la vorticosità di confine devono cadere a zero in corrispondenza delle estremità delle ali. Lo strato vorticoso *(a sinistra)* è costituito da filetti, o vortici infinitesimi allineati approssimativamente nella direzione del volo e di intensità crescente verso le estremità delle ali, dove si accumulano per formare i vortici di estremità. Il campo collettivo di velocità dei filetti vorticosi impartisce una «deflessione», o velocità indotta, allo strato vorticoso come parte del processo di accumulo. Come mostrato a destra, la velocità di deflessione *(w)* si somma vettorialmente con la velocità di volo nello spazio *(V)* per produrre una velocità locale *(W)* disallineata rispetto alla velocità di volo di un angolo pari all'angolo indotto di attacco *(α)*. Per produrre un certo valore di portanza l'angolo geometrico di attacco della sezione alare deve essere maggiore dell'angolo effettivo di attacco della superficie aerodinamica di un valore uguale all'angolo indotto. Quindi il vettore di portanza *(dL/dy)* risulta ruotato all'indietro di un angolo uguale all'angolo indotto, dando luogo a una resistenza aerodinamica indotta.

La distribuzione ottimale della portanza, corrispondente alla minima resistenza aerodinamica indotta, si ottiene quando la portanza e la circolazione «al limite» variano secondo una funzione ellittica fra le due estremità delle ali. Con tale distribuzione, il campo di velocità indotto dal sistema vorticoso sul bordo di uscita spinge lo strato vorticoso verso il basso con una velocità indotta di deflessione uniforme su tutta l'apertura alare. In queste condizioni per un dato valore di portanza viene trasferita alla scia vorticosa una minima quantità di energia cinetica.

La teoria dei vortici per le eliche è analoga alla teoria dei vortici per le ali, salvo per il fatto che gli strati vorticosi sul bordo di uscita sono superfici elicoidali *(in alto)*. Gli strati si accumulano in tanti vortici di estremità per quante sono le pale e in un vortice centrale la cui energia è uguale alla somma delle energie dei vortici di estremità. I vettori di velocità indotta *(w)* risultano perpendicolari alla superficie degli strati elicoidali e corrispondono in ampiezza al valore dello scorrimento nella teoria delle quantità di moto di Rankine-Froude, valida per le eliche marine. Allo scopo di avere la minima perdita indotta, ogni filetto vorticoso elicoidale *(in basso)* si dovrebbe muovere in direzione perpendicolare al proprio asse in modo tale che l'intero strato vorticoso sia in grado di muoversi assialmente con una velocità di «spostamento» costante. Il movimento dello strato vorticoso per ottenere la perdita minima venne scoperto da un collega di Prandtl, Albert Betz.

viscoso ideale soddisfino il teorema di Bernoulli, non può esservi spinta senza circolazione. Ne consegue che la spiegazione abituale fornita nelle scuole secondarie per la portanza delle superfici aerodinamiche in base al teorema di Bernoulli è del tutto inadeguata.

Secondo quella spiegazione la spinta si determina perché l'aria che scorre sulla superficie curva superiore di un tipico profilo aerodinamico deve compiere un percorso più lungo di quella che scorre sotto la superficie piatta inferiore ed è quindi costretta a muoversi con velocità maggiore. Perciò la pressione sulla faccia superiore dell'ala deve essere minore rispetto a quella sulla faccia inferiore, con una conseguente spinta verso l'alto; imprime cioè una portanza. In realtà la portanza può prodursi anche su un profilo aerodinamico perfettamente simmetrico, inclinato verso l'alto di un opportuno angolo di attacco. A questo punto sorge spontanea la domanda di come la corrente d'aria sappia dividersi in corrispondenza del bordo di attacco e riunirsi al bordo di uscita in modo che il flusso al di sopra del dorso dell'ala abbia in media una velocità maggiore e una pressione minore del flusso sotto il ventre dell'ala stessa.

L'angolo di attacco è ovviamente decisivo, ma non è disponibile alcuna teoria fisica soddisfacente che spieghi perché la corrente si divide esattamente come avviene in realtà. La regolazione di Joukowsky dell'intensità del vortice per produrre il ristagno nel punto singolare del cilindro trasformato (cioè del profilo aerodinamico) ha l'effetto di far coincidere il campo di flusso con il bordo di uscita del profilo aerodinamico, una supposizione plausibile definita ipotesi di Joukowsky o ipotesi di Kutta-Joukowsky. L'effetto di far coincidere il flusso in corrispondenza del bordo di uscita è di disallineare, ovvero sfalsare, il flusso nel punto ove si divide in corrispondenza del bordo di attacco nel profilo aerodinamico in modo che ne risultino circolazione e spinta. Il bordo di attacco dovrebbe essere opportunamente arrotondato, come lo sono i corpi matematici a profilo aerodinamico di Joukowsky.

L'angolo di attacco di un profilo aerodinamico può essere modificato inclinando l'asse che congiunge i punti singolari anteriore e posteriore della trasformazione rispetto alla direzione del campo di flusso remoto, cioè rispetto alla direzione del campo prima che venga influenzato dal profilo aerodinamico. La curvatura del profilo aerodinamico può essere variata sfalsando il raggio congiungente il centro del cilindro con il punto singolare posteriore rispetto all'asse dei punti singolari nella trasformazione. Ciascuno di questi disallineamenti implica un cambiamento nella resistenza della circolazione, o vertice di confine, e dà luogo a corrispondenti cambiamenti nella portanza. Le prove nella galleria del vento dei profili aerodinamici di Joukowsky eseguite da Joukowsky in Russia e da Prandtl in Germania, entrambi con i relativi colleghi, poco prima dell'inizio della prima guerra mondiale, consentirono di verificare l'esattezza della teoria per piccoli angoli di attacco e per piccole curvature. Tutte le moderne teorie aerodinamiche comprendono le conclusioni di Joukowsky come caso particolare.

La teoria di Joukowsky, benché spieghi la portanza di sezioni d'ala in un flusso bidimensionale (cioè in assenza di componenti lungo l'apertura alare nella velocità del flusso), presenta difficoltà di applicazione nel caso di ali con apertura finita, nelle quali esiste una componente in direzione dell'apertura alare e dove la portanza e il vortice al limite devono ridursi gradualmente a zero in corrispondenza delle estremità delle ali. Il problema venne risolto a Göttingen nel periodo immediatamente precedente e durante la prima guerra mondiale da Prandtl, Betz, Max Munk e C. Wieselberger, i quali introdussero il concetto di uno strato vorticoso sul bordo di uscita costituito da vor-

La scia vorticosa di un'elica aerea a due pale *(in alto)* è stata fotografata in una galleria del vento alla fine degli anni cinquanta da F.M.N. Brown dell'Università di Notre Dame. La fotografia mostra come i filamenti di fumo facenti parte di una schiera verticale vengano attorcigliati in una serie di vortici di estremità distanziati uniformemente dal passo dell'elica. Il vortice in corrispondenza del mozzo è spezzato dalla custodia dell'albero motore. Il diagramma mostra come le discontinuità curvilinee corrispondano alle sezioni lungo gli strati vorticosi diffusi dalle pale.

tici lineari infinitesimali approssimativamente allineati secondo la direzione del volo. L'intensità di ciascuno di questi vortici è proporzionale al gradiente locale in direzione dell'apertura alare della vorticosità al limite, associata con il valore locale di portanza della sezione aerodinamica sul modello di Joukowsky.

I vortici sul bordo di uscita sono dotati di particolare energia in prossimità dei confini dello strato alle estremità alari ove il gradiente tende all'infinito. Lo strato vorticoso sul bordo di uscita si accumula intorno ai suoi intensi orli esterni, formando due distinti «vortici di estremità» (dei quali tuttavia non si era tenuto conto nella teoria originaria). Il processo di accumulazione risulta sostanzialmente completo dietro il bordo di uscita a una distanza pari a diverse volte l'apertura alare, dove i vortici di estremità sono separati da una distanza inferiore alla medesima apertura alare. Quando l'umidità negli strati superiori dell'atmosfera si trova entro certi valori, il vapor d'acqua contenuto nei gas di scarico dei motori di un aereo condensa all'interno dei nuclei dei vortici di estremità rendendoli visibili sotto forma di scie di condensazione.

Il campo collettivo di velocità dei vortici all'interno dello strato vorticoso origina una «deflessione», o velocità indotta, che tende a trasportare lo strato verso il basso in ogni stadio del processo di accumulazione. In particolare se la portanza dell'ala è distribuita secondo una funzione ellittica lungo l'apertura alare, lo strato vorticoso (prima di accumularsi) viene trasportato verso il basso con velocità uniforme per tutta la sua larghezza. L'energia cinetica associata con il campo di flusso vorticoso sul bordo di uscita non è recuperabile, ma può essere minimizzata per certi valori di portanza, di velocità di volo e di apertura alare se la portanza è distribuita ellitticamente. Tale importante risultato è attribuito a Munk.

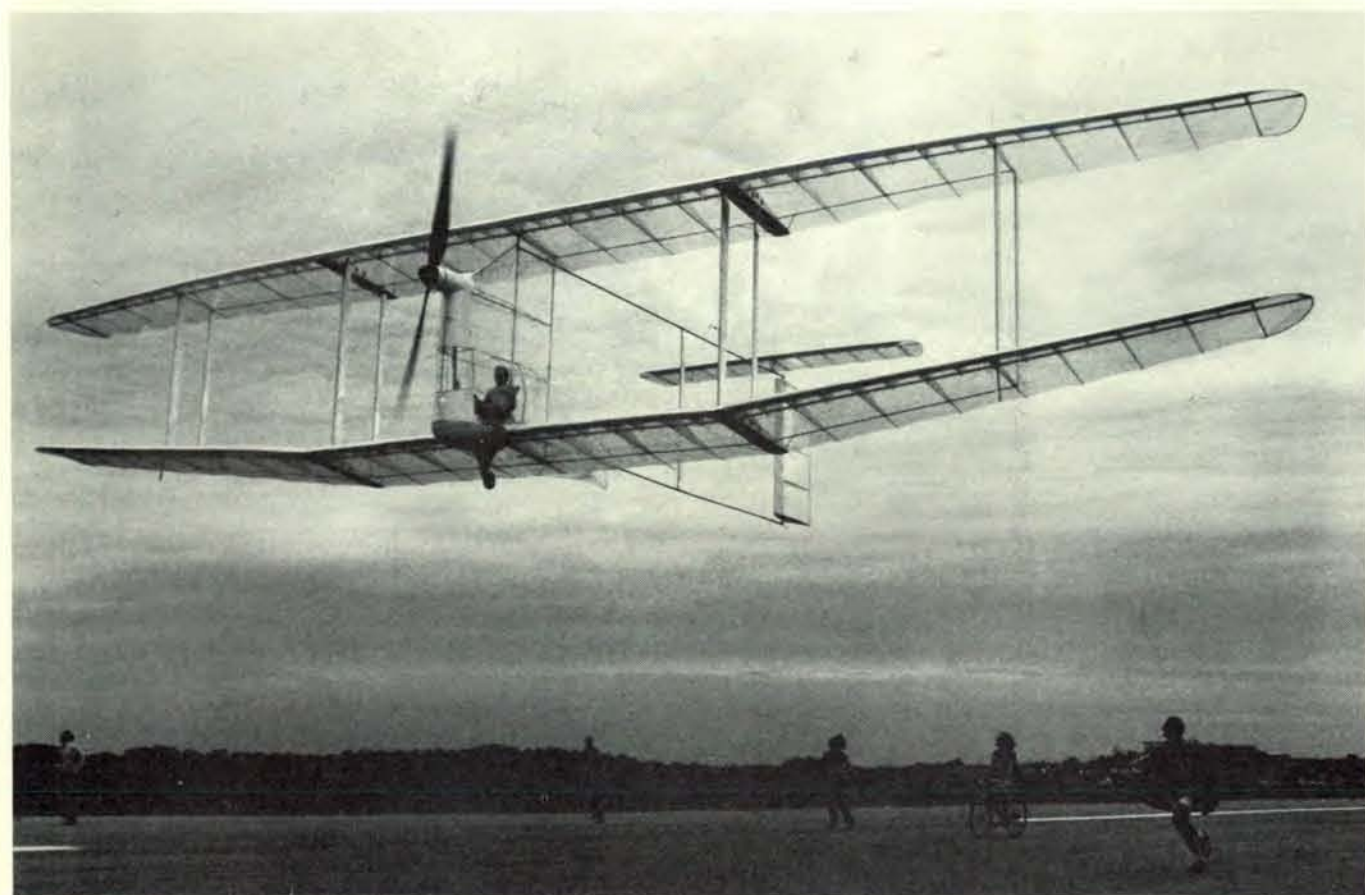
Il modello di Joukowsky per la portanza aerodinamica in un flusso bidimensionale può essere incorporato nella teoria dei vortici per ali di apertura finita, calcolando l'ampiezza della deflessione indotta dal sistema vorticoso sul bordo di uscita. La somma vettoriale delle velocità di deflessione e di volo determina un angolo indotto di attacco. Il necessario angolo geometrico di attacco rispetto all'area non perturbata si ottiene sommando l'angolo indotto di attacco all'angolo teorico di attacco richiesto per un'ala di apertura infinita in un flusso bidimensionale (di Joukowsky).

Anche i vettori della portanza alare vengono ruotati all'indietro di un angolo uguale all'angolo indotto di attacco dando luogo pertanto a una resistenza aerodinamica indotta. Infine il tasso con cui viene perduta l'energia, corrispondente al prodotto della resistenza aerodinamica indotta per la velocità di volo, compensa esattamente l'energia cinetica erogata con continuità al sistema di vortici. Vi è inoltre una resistenza di attrito sulle sezioni alari dovuta all'azione della viscosità negli strati limite adiacenti alle superfici alari, un effetto a suo tempo già descritto da Prandtl. Le previsioni della teoria dei vortici furono confermate da esperimenti nella galleria del vento a Göttingen, durante la prima guerra mondiale.

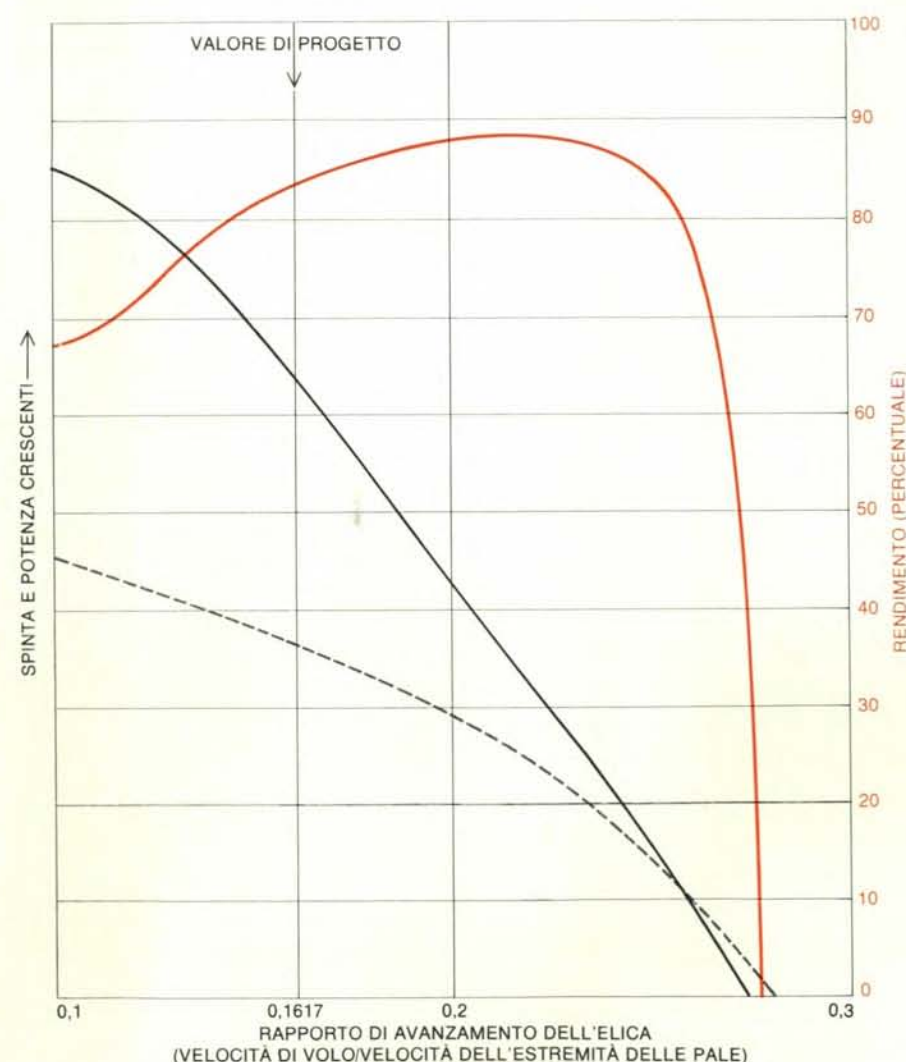
Un breve passo separa questo stadio dello sviluppo dalla teoria dei vortici per le eliche. La differenza essenziale risiede nel fatto che gli strati vorticosi di uscita di un'elica sono all'inizio elicoidali. Alla fine tali strati si accumulano in tanti vortici di estremità quante sono le pale e in un vortice centrale (o vortice di mozzo) rotante in senso inverso ai primi e la cui intensità è uguale alla somma delle intensità dei singoli vortici di estremità. Il campo collettivo di velocità degli strati vorticosi elicoidali è identico alla scia dell'elica rotante. E proprio come esiste un solo moto dello strato vorticoso che rende minime le perdite indotte dell'ala portante, così vi è un solo moto dello strato vorticoso che minimizza le perdite indotte di un'elica traente. Questo moto venne scoperto da Betz, mentre la corrispondente (ma approssimata) distribuzione radiale della vorticosità al limite per un modello di Joukowsky con distribuzione ottimale della spinta delle pale fu calcolata da Prandtl. I due ricercatori pubblicarono i risultati cui erano pervenuti nel 1919. Goldstein entrò a far parte del gruppo negli anni venti e determinò gli errori nella distribuzione approssimata della circolazione di Prandtl nella sua tesi di dottorato nel 1929.

È forse giusto accennare che perfino gli ingegneri aeronautici hanno incontrato difficoltà nell'impadronirsi completamente dei contributi di Betz-Prandtl-Goldstein perché gli strati vorticosi di un'elica e il loro sviluppo geometrico nello spazio sono di difficile visualizzazione tridimensionale. Dopo tutto un'elica altro non è che un tipo particolare di ala, ma al contrario di questa, che in volo rettilineo incontra l'aria alla stessa velocità media in ogni sezione e quindi ha il medesimo angolo di attacco dalla radice all'estremità, le varie sezioni dell'elica incontrano l'aria a velocità che variano dal mozzo alla punta. Questo è il motivo per cui le pale delle eliche sono svergolate. In prossimità del mozzo, dove la velocità assiale o di avanzamento della sezione di pala è elevata rispetto alla velocità periferica, la corda, o campanatura, della pala si scosta solo di poco dalla direzione di volo. All'estremità della pala, ove la velocità periferica è elevata rispetto alla velocità di avanzamento, la corda può essere quasi ad angolo retto con la direzione di volo.

Come si è visto, il campo di velocità dello strato vorticoso di uscita crea una deflessione perpendicolare alla direzione del moto del profilo aerodinamico nello spazio non perturbato. In prossimità del mozzo dell'elica la deflessione è quindi volta solo leggermente all'indietro rispetto alla linea di volo. In altre parole, il vettore della velocità indotta vicino al mozzo ha un'elevata componente periferica e solo una piccola componente assiale. Di conseguenza le sezioni delle pale vicine al mozzo contribuiscono in minima parte alla spinta complessiva dell'elica. Al contrario, all'estremità delle pale la velocità indotta ha un'elevata componente assiale e una componente periferica molto minore. Ne consegue che la spinta per unità di raggio di una pala a sezione costante aumenta dal mozzo all'estremità, tranne che all'estremità stessa, dove deve ridursi a zero «ellitticamente».

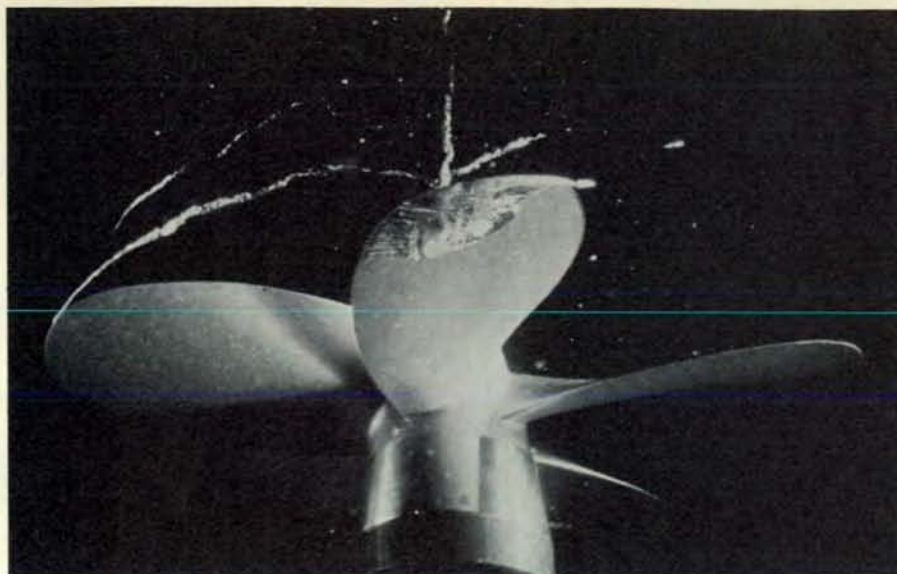
Si potrebbe pensare che, poiché la faccia, o parte esterna della pala, compie un lavoro utile molto maggiore del dorso o parte interna, nel progetto di un'elica la maggiore attenzione debba essere dedicata alla faccia. Si potrebbe immaginare, per esempio, un'elica le cui pale aumentassero di larghezza dal mozzo verso le estremità; una geometria del genere (anche se realizzabile dal punto di vista strutturale) avrebbe un rendimento relativamente basso. Per l'elica l'equivalente della velocità uniforme di deflessione lungo l'apertura alare per un'ala ad alto rendimento si realizza quando la velocità indotta in corrispondenza di ogni sezione della pala dell'elica è pari a metà della velocità dello strato vorticoso in corrispondenza della stessa sezione e quando le velocità assiale e periferica dello strato

Il biplano *Chrysalis*, azionato a pedali, è stato costruito dagli studenti del Massachusetts Institute of Technology fra marzo e giugno dello scorso anno. La configurazione a biplano è stata scelta per minimizzare la resistenza aerodinamica. L'apertura alare è stata limitata a 22 metri per mancanza di maggiore spazio nell'aviorimessa. A causa sia dell'apertura alare relativamente piccola sia del peso di 43 chilogrammi (30 per cento di più del *Gossamer Albatross*), il *Chrysalis* richiedeva una potenza ai pedali compresa almeno fra 0,3 e 0,4 cavalli-vapore, a seconda del peso del pilota. L'elica del *Chrysalis*, analogamente a quella del *Gossamer Albatross*, è stata progettata al calcolatore secondo un programma elaborato dallo studente Hyong Bang sulla base di un algoritmo dell'autore. La fotografia è stata scattata da Steven Finberg.

Il *Gossamer Albatross* è stato ripreso in fase di avvicinamento alla costa francese nel giugno 1979, mentre gareggiava per un premio di 100 000 sterline offerto da Henry Kremer, un industriale britannico. Allen, il pilota, è un ciclista professionista di 27 anni, del peso di 62 chilogrammi, e appassionato di volo a vela. Il volo di 35,82 chilometri, ostacolato da vento di prua, ha richiesto 2 ore e 49 minuti, alla velocità di 12,7 chilometri all'ora. Il velivolo non è mai salito a più di 4,5 metri dal pelo dell'acqua. La strumentazione di cui disponeva Allen era costituita da un anemometro e da un altimetro che incorporava un telemetro a sonar ricavato da una macchina fotografica Polaroid SX-70. La velocità media di pedalata è stata di 75 giri al minuto, corrispondente a 95 giri al minuto dell'elica. La fotografia è stata scattata da Don Monroe.

Vista frontale *(in alto)* e laterale *(in basso)* dell'elica del *Chrysalis*. La «forma sviluppata in pianta», indicata con le linee tratteggiate, è la geometria che avrebbe la pala se non fosse svergolata. La corda massima è al 30 per cento del raggio. L'elica ha un diametro di 4,27 metri, maggiore del 4 per cento di quello dell'elica del *Gossamer Albatross*. Entrambe le eliche sono state progettate in base a un algoritmo dell'autore per la minima perdita indotta. Il *Chrysalis* è stato progettato per un «motore» di 0,5 HP (373 watt) a 135 giri al minuto dell'elica, una velocità di volo di 4,88 metri al secondo (17,54 chilometri all'ora) e quindi per un «rapporto di avanzamento» (la velocità di volo divisa per la velocità dell'estremità delle pale) di 0,1617. Le eliche di questo tipo, a causa dell'elevata velocità delle estremità delle pale (6,26 volte la velocità di volo), non sarebbero adatte per aeroplani con velocità di volo superiore a 180 chilometri all'ora.

Le prestazioni dell'elica del *Chrysalis* mostrano come i coefficienti di spinta *(curva continua in nero)*, di potenza *(curva tratteggiata)* e di rendimento complessivo *(in colore)* varino con il rapporto di avanzamento dell'elica. La minima perdita indotta per il carico sul disco battuto di progetto si verifica quando il rendimento dell'elica si avvicina all'83 per cento. Il rendimento cresce con maggiori rapporti di avanzamento (ciò equivale a una rotazione più lenta dell'elica a parità di velocità di volo) anche se le perdite indotte non sono più al valore minimo; queste diventano minori riducendo il carico sull'elica e diminuendo la velocità assiale media della corrente di scorrimento. I coefficienti di spinta e di potenza sono maggiori quando l'elica ruota più velocemente del valore di progetto *(curve a sinistra del valore di progetto)* e peggiorano a minor numero di giri.

vorticoso variano con il raggio in modo tale da presentare l'aspetto di una velocità assiale di «spostamento» uniforme. Un'insegna rotante con strisce elicoidali ha, per esempio, una velocità assiale di spostamento finita, ma non una effettiva velocità assiale fisica.

In generale la geometria di un'elica a massimo rendimento, denominata geometria delle minime perdite indotte, è quella di un'elica le cui pale hanno una corda massima compresa fra il 30 e il 40 per cento del raggio e sono rastremate verso l'estremità realizzata a punta. Queste pale erano di impiego abituale con le eliche in lega di alluminio costruite fra la fine degli anni venti e l'inizio degli anni trenta. Con lo sviluppo di motori sovralimentati e di maggiore potenza impiegati per raggiungere quote più elevate divenne abituale aumentare la superficie delle pale verso l'estremità al fine di assorbire la maggior potenza senza aumentare il diametro dell'elica. Gli aumenti di diametro erano dichiarati inammissibili in quanto le estremità delle pale raggiungevano già velocità transoniche, velocità cioè all'incirca pari a quella del suono, un regime per il quale non si conoscevano profili di pala soddisfacenti. Nei casi limite le estremità venivano realizzate perfino in forma quadrata. Tali scostamenti dalla geometria di minima perdita indotta davano luogo a piccole riduzioni nel rendimento delle eliche, che non erano avvertite dalla grande maggioranza degli ingegneri aeronautici di quel tempo, moltissimi dei quali non avevano probabilmente letto, o forse non avevano compreso, il lavoro di Goldstein (o i commenti di Hermann Glauert e più tardi di Theodore Theodorsen). Il profilo di minime perdite indotte è di adozione essenziale solo quando il progettista ha bisogno di disporre di qualsiasi sia pur minima frazione di rendimento.

Il mio contributo alla teoria delle eliche è consistito nello sviluppo di una teoria della quantità di moto a «gradiente radiale», adatta al trattamento al calcolatore e compatibile con l'enunciato di Betz-Prandtl della teoria dei vortici per le eliche. Tutte le teorie delle eliche cercano di calcolare la velocità indotta degli elementi delle pale dovuta al sistema dei vortici sul bordo di uscita. Nel caso di eliche con minima perdita indotta la velocità a metà spostamento corrisponde alla velocità di scorrimento nella teoria del disco attuatore di Rankine e Froude. Come nella teoria delle ali le componenti assiale e periferica della velocità indotta (denominata velocità di afflusso nella teoria delle eliche) richiedono un aumento nell'angolo delle pale per raggiungere uno specificato valore di spinta in unione con una rotazione all'indietro dei vettori di scorrimento delle pale, con un conseguente incremento della coppia necessaria per far girare l'elica. La teoria dei vortici per le eliche è pertanto affatto compatibile con le prime teorie della quantità di moto relative alle eliche marine e giustifica i dettagli del carico delle pale corrisponden-

Le eliche marine in movimento vengono studiate al laboratorio di idrodinamica navale del MIT con l'ausilio della fotografia ad alta velocità da membri del Department of Ocean Engineering. In questa fotografia, che mostra l'elica vista dall'alto, la regione bianca all'estremità della pala di mezzo è una grossa bolla di vapor d'acqua, che evidenzia un'improvvisa cavitazione prodotta da un campo di flusso non uniforme simulante la scia dello scafo di una nave. Le basse pressioni che si formano sulla faccia, cioè sulla superficie «aspirante» delle pale provoca l'ebollizione istantanea dell'acqua circostante. La rotazione della pala frantuma la grossa bolla che si suddivide in un flusso di bolle più piccole. Una cavitazione ciclica provoca la riduzione del rendimento dell'elica, con conseguenti vibrazioni e genera una caratteristica «sigla» acustica. La cavitazione può essere ridotta aumentando la superficie della pala e ottimizzando la geometria della stessa.

te a una particolare geometria dell'elica.

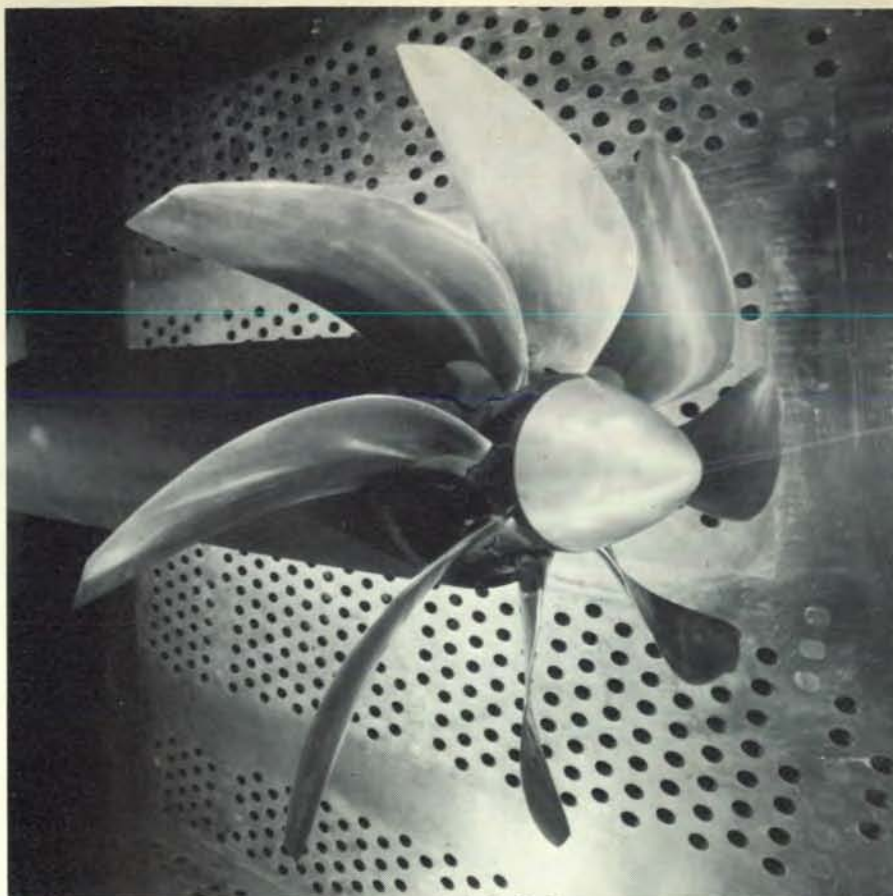
Le eliche, come del resto le ali, hanno anche perdite di «profilo» o di attrito, dovute all'azione della viscosità dell'aria che scorre in prossimità del profilo aerodinamico, cioè dell'aria contenuta nello strato limite. Queste perdite vengono minimizzate facendo lavorare gli elementi delle pale con angoli di attacco per i quali i rapporti fra la resistenza aerodinamica e lo scorrimento sono bassi e, se possibile, scegliendo angoli del passo elicoidale pari a 45 gradi meno metà dell'angolo la cui tangente è il rapporto fra resistenza aerodinamica e scorrimento della pala bidimensionale. La riduzione delle perdite di attrito richiede quindi «rapporti di avanzamento» abbastanza elevati (rapporti fra la velocità di volo e la velocità dell'estremità delle pale pari a circa 0,7) accompagnati da un numero di pale esiguo, in modo che la solidità (ovvero il rapporto fra la superficie delle pale e quella del disco battuto) possa essere concentrata in poche pale di buona larghezza funzionanti a numeri di Reynolds elevati, con corrispondente basso attrito superficiale (il numero di Reynolds è un rapporto fra le forze di coppia e le forze di attrito in un fluido e aiuta a determinare il carattere dello strato limite e la sua resistenza alla separazione). D'altra parte le perdite indotte vengono minimizzate aumentando il numero delle pale e abbassando il rapporto di avanzamento al fine di diminuire la distanza fra i singoli strati vorticosi e di renderli quasi perpendicolari all'asse dell'elica. L'obiettivo è di fare in modo che il campo della velocità di scia risulti quanto più uniforme possibile. Le caratteristiche che consentono di tenere basse le perdite indotte e le perdite di attrito sono pertanto in contrasto fra loro: eliche con rendimenti massimi richiedono un equilibrio ottimale fra le due.

Usando la distribuzione analitica approssimata della circolazione di Betz-Prandtl, sono riuscito a definire un algoritmo, cioè un procedimento di calcolo, che consente al progettista di stabilire la geometria di un'elica con perdita indotta minima, dopo aver accettato qualsivoglia vincolo di progetto che ponga un limite inferiore nel raggiungere la minore perdita indotta possibile (per esempio se il progetto richiede per una qualsiasi ragione un'elica a due pale il progettista non può esaminare i possibili vantaggi dovuti all'aggiunta di altre pale). Questo algoritmo permette di progettare un'elica che sviluppi una spinta data o che assorba una potenza assegnata con il maggior rendimento possibile e con diametro, numero di pale, velocità di volo, numero di giri dell'albero e densità dell'aria prefissati. L'algoritmo tiene conto anche di un ampio ventaglio di rapporti fra portanza e resistenza aerodinamica per elementi di pala con diversi profili al fine di ottenere i voluti coefficienti di portanza in corrispondenza di particolari circostanze di progetto. Tale algoritmo, poiché fornisce il rendimento dell'elica non solo nella condizione di progetto, ma anche in condizioni di volo molto discoste dalla condizione di progetto, può essere impiegato dal progettista per cambiare la condizione di progetto medesimo.

È stata una fortuna per me partecipare a una dimostrazione pratica della validità di quella che potrebbe essere propriamente definita una teoria classica dell'aerodinamica delle eliche, pubblicata per la prima volta nel 1919 (un anno prima della mia nascita) e in pratica dimenticata fino al 1942; in tale anno iniziai la carriera professionale nella fabbrica di Buffalo, N. Y., della Curtiss-Wright Corporation. Sono debitore di questa dimostrazione a un gruppo di studenti entusiasti del Massachusetts Institute of Technology (MIT), che volevano costruire nel gennaio 1979 un aereo azionato a pedali e che mi richiesero le nozioni per il progetto delle eliche, nozioni raccolte in previsione di una presentazione a un simposio.

Gli studenti erano stimolati dal sogno (di certo ben lontano) di vincere il premio di 100 000 sterline offerto da Henry Kremer, un industriale inglese, per un aeroplano mosso dall'uomo e capace di sorvolare il Canale della Manica. (Il *Gossamer Condor,* un velivolo ultraleggero, progettato da Paul MacCready presidente della Aero Vironment Incorporated di Pasadena, California, aveva vinto nel 1977 un precedente premio Kremer di 50 000 sterline per il primo aeroplano mosso dall'uomo in grado di percorrere un intero percorso a forma di otto intorno a due piloni distanti 800 metri.)

Come primo passo gli studenti Hyong Bang, Robert Parks e Harold Youngren codificarono una versione dell'algoritmo delle eliche per consentire il calcolo a macchina e l'applicazione per la progettazione dell'elica di un modello radiocomandato di biplano in scala 1 a 8, il cui nome avrebbe dovuto essere *Chrysalis.* Il modello, che aveva un'apertura alare di 2,74 metri, era azionato dal motore di una macchina fotografica Polaroid tipo SX-70, che faceva ruotare l'elica tramite un ingranaggio con rapporto di riduzione de 11 a 1. Il modello venne costruito nel corso di una settimana, in febbraio. Mentre un'elica progettata su basi intuitive avrebbe consentito al modello di mantenersi a malapena in volo con batterie a piena carica, l'elica teoricamente corretta corrispondeva perfettamente alle esigenze della cellula e della potenza del motore, consentendo al modello di compiere tre volte e mezzo un percorso a forma di otto intorno a due piloni distanti 20 metri nella palestra del MIT prima che le batterie si scaricassero.

Il secondo successo per la resuscitata teoria di Betz-Prandtl venne quando gli stessi studenti progettarono una nuova elica per il monoplano *Gossamer Albatross,* che il gruppo di MacCready aveva costruito per partecipare al premio della traversata della Manica. Bryan Allen, un giovane ciclista professionista che pesava 62 chilogrammi, era stato capace di mantenere in aria il *Gossamer Albatross* dotato dell'elica originale per appena 17 minuti. Con la nuova elica del MIT, assieme ad altri miglioramenti nella struttura del velivolo, Allen riteneva di poter mantenere il velivolo in aria per alcune ore o almeno fin quando la sua «riserva di car-

Il *propfan* transonico è uno dei diversi progetti in corso di studio presso la United Technologies Corporation e la National Aeronautics and Space Administration per un potenziale impiego in aeroplani a turboelica capaci di raggiungere la velocità degli aerei di linea a reazione. Se il nuovo progetto avrà successo, le macchine che ne saranno equipaggiate consumeranno circa il 20 per cento di carburante in meno dei tipi attuali. Il *propfan* è denominato transonico perché la velocità delle estremità delle pale supera la velocità del suono del 13 per cento circa. Il velivolo volerà all'80 per cento della velocità del suono (circa 980 chilometri all'ora a una quota di 10 600 metri). Per poter funzionare in regime transonico, un'elica aerea deve avvicinarsi alla elevata solidità delle eliche marine. La solidità di quest'elica è del 70 per cento. Le pale larghe forniscono gli elevati rapporti fra portanza e resistenza aerodinamica occorrenti per il funzionamento transonico. Le forme curve in pianta delle pale sono appositamente scelte per ridurre il rumore «di prossimità» (interno della cabina) causato dalla focalizzazione del campo di pressione rotante.

burante» (costituita da glucosio e acqua) si fosse esaurita. La traversata effettiva della Manica, ostacolata da vento di prua, richiese ad Allen di pedalare per 2 ore e 49 minuti. La distanza coperta fu di 35,82 chilometri con una velocità media rispetto al suolo di 12,7 chilometri all'ora.

Il terzo successo per la teoria di Betz-Prandtl venne con il progetto dell'elica per il vero *Chrysalis*. Costruito fra marzo e giugno, il biplano fece più di 320 voli nel corso dell'estate, prima di essere definitivamente smontato in settembre. Se il *Chrysalis* non avesse avuto un'elica appropriata dal punto di vista teorico e perfettamente adattata alla struttura e al propulsore (0,5 HP) è dubbio che, con la piccola apertura alare e il notevole peso (10 chilogrammi di più del *Gossamer Albatross*) di cui era dotato, avrebbe addirittura potuto alzarsi da terra.

I problemi dei progettisti di eliche marine sono ben più vincolanti di quelli dei progettisti di eliche aeree. Il diametro del disco battuto è limitato da una parte dal massimo pescaggio consentito per la nave a pieno carico e dall'altra dalla linea di galleggiamento quando la nave è scarica. L'elica si trova in un campo di flusso fortemente disuniforme, distorto dal profilo della poppa e con moltissime irregolarità di velocità prodotte dagli strati limite intorno allo scafo. Malgrado una favorevole geometria a pale multiple, la decisa limitazione nel diametro dà luogo a un carico eccessivo sul disco battuto e a forti perdite indotte. La solidità dell'elica, già elevata a causa del forte carico sul disco battuto, viene ancora aumentata dai vincoli imposti sull'angolo di attacco delle pale e sulla loro freccia per la necessità di limitare il fenomeno della cavitazione. La cavitazione si verifica quando la velocità del fluido in prossimità delle pale diventa tanto elevata che la pressione locale scende al di sotto della pressione di vapore dell'acqua. Ciò determina l'ebollizione del liquido, con la creazione improvvisa di cavità nel campo di flusso. La cavitazione è la principale causa di rumore nelle eliche marine.

Ulteriori problemi sono presentati dalle frequenze di risonanza per eccitazione che possono istituirsi fra la pala rotante e lo scafo della nave. Le eliche marine in genere sono dotate di pale ricurve con angoli di attacco a freccia positiva in modo da rendere l'eccitazione una funzione meno marcata dell'angolo di rotazione delle pale. Per tutte queste ragioni l'analisi delle eliche marine deve tener conto sia delle variazioni nella velocità di afflusso lungo tutta la superficie della pala sia dell'accumulo dello strato vorticoso, che si instaura molto rapidamente nel caso di eliche fortemente caricate. La semplice distribuzione approssimata della velocità di afflusso e della circolazione di Betz, Prandtl e Goldstein non è applicabile alle eliche marine: il progettista deve ripiegare su metodi numerici semiempirici di analisi. Infine il rendimento delle eliche marine è di solito intorno al 70 per cento. Gli architetti navali accettano questo risultato come il migliore ottenibile, dati i pesanti vincoli pratici.

Al momento attuale viene dato ampio spazio agli studi sulle eliche per aerei la cui velocità raggiunge l'80 per cento della velocità del suono, corrispondente a circa 850 chilometri all'ora, ad altezze di crociera comprese fra 10 700 e 13 700 metri. Le estremità delle pale di queste eliche raggiungono anch'esse velocità periferiche pari all'80 per cento della velocità del suono, e quando viene aggiunta la velocità di volo, la velocità dell'estremità della pala raggiunge il 113 per cento della velocità del suono. A tali velocità le eliche devono essere multipala, e quindi con elevata solidità, perché possano presentare buoni rapporti fra portanza e resistenza aerodinamica e pertanto assomigliano da vicino alle eliche marine. Si dovrebbero poter raggiungere rendimenti dell'80 per cento. Le eliche ad alta velocità avranno un funzionamento rumoroso (per ragioni del tutto differenti da quelle che rendono rumorose le eliche marine). Tuttavia il rumore sarà in prevalenza un rumore di «prossimità» che può essere reso tollerabile migliorando l'insonorizzazione della fusoliera. Il rumore «a distanza» in realtà può essere inferiore a quello dei motori a reazione esistenti. Il rendimento termico di un motore a turboelica, purché equipaggiato con un'elica adatta per le alte velocità, dovrebbe essere del 20 per cento superiore al rendimento di un motore a reazione a forte by-pass dotato di un «generatore a gas» (la combinazione di un compressore, una camera di combustione e una turbina motrice) avente la stessa dimensione di quella occorrente per il motore a turboelica. L'elica del motore a turbina dovrebbe avere un diametro circa doppio del compressore nel motore a reazione.

Dopo oltre un secolo di applicazioni navali e aeronautiche l'elica è ben viva e vitale: c'è da attendersi che essa sopravviverà fin quando l'uomo impiegherà l'acqua e l'aria per navigare.

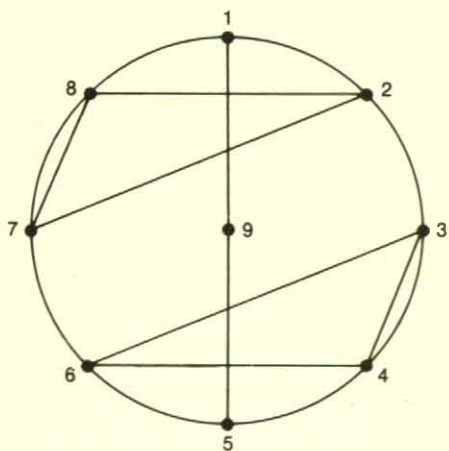
GIOCHI MATEMATICI

di Martin Gardner

Che cosa hanno in comune invitati a pranzo, studentesse in passeggiata e prigionieri ammanettati?

Una donna decide di invitare a cena 15 amici. Vuole avere a cena, per 35 giorni, esattamente tre amici al giorno e vuole ordinare le triple in modo tale che ogni coppia di amici venga solo una volta. È possibile la cosa?

Questo e altri problemi analoghi, che appartengono a una vasta area della matematica combinatoria chiamata «block-design theory», vennero analizzati intensamente nel XIX secolo soprattutto come problemi ricreativi. Più tardi ci si accorse che avevano un ruolo importante in statistica, particolarmente nella predisposizione di esperimenti scientifici. Una piccola branca della block-design theory si occupa dei sistemi di triple di Steiner, di cui il problema degli ospiti a cena costituisce un semplice esempio. Jakob Steiner, uno studioso di geometria svizzero, fu il pioniere dello studio di questi sistemi nel XIX secolo.

In generale, un sistema di triple di Steiner è una disposizione di n oggetti in triple in modo che ogni coppia di oggetti appaia in una tripla una e una sola volta. È facile dimostrare che il numero di coppie è $1/2n(n-1)$ e che il numero di triple richieste è un terzo del numero di coppie, ossia $1/6n(n-1)$. Naturalmente, un sistema di triple di Steiner è possibile solo quando entrambi questi numeri sono interi. Questo avviene quando n è congruo a 1 o 3 modulo 6, vale a dire c'è un resto pari a 1 o 3 quando n è diviso per 6. Perciò la sequenza di possibili valori di n è 3, 7, 9, 13, 15, 19, 21 e così via.

Con soli tre ospiti, il problema della cena ha una soluzione banale: vengono tutti lo stesso giorno. Dato che le triple di Steiner non sono ordinate, la soluzione è naturalmente unica. C'è un'unica soluzione anche per sette ospiti: (1, 2, 4), (2, 3, 5), (3, 4, 6), (4, 5, 7), (5, 6, 1), (6, 7, 2) e (7, 1, 3). L'ordine delle triple e l'ordine dei numeri in ogni tripla può essere modificato a piacere: basta non cambiare lo schema di fondo. In più, i numeri possono anche essere scambiati. Per capire questo punto si immagini che ogni ospite abbia un bottone con sopra dipinto un numero. Se due o più ospiti scambiano i propri bottoni, si può considerare la nuova combinazione uguale a quella precedente.

Analogamente, per nove ospiti c'è un'unica soluzione, per 13 ospiti ci sono due soluzioni e per 15 ospiti si sa da tempo che ci sono 80 soluzioni base. Per valori di n maggiori di 15, il numero di soluzioni distinte non è noto, anche se è stato dimostrato che c'è una soluzione per ogni valore di n. Per $n = 19$ ci sono centinaia di migliaia di soluzioni.

Proponiamo ora una lieve complicazione dei sistemi di triple di Steiner per renderle più interessanti. Supponiamo che la donna decida di invitare tutti e 15 gli amici ogni giorno per sette giorni, facendoli sedere a tre a tre a cinque tavoli e facendo in modo che ogni coppia di amici stia assieme a tavola una sola volta.

Il nostro nuovo problema è equivalente a uno dei più famosi rompicapo della storia della matematica combinatoria: il problema delle studentesse di Kirkman, così chiamato dal nome del reverendo Thomas Penyngton Kirkman, un matematico inglese dilettante del XIX secolo che fu rettore della chiesa di Croft, nel Lancashire, per più di 50 anni. Sebbene fosse completamente autodidatta in matematica, le sue scoperte furono così numerose e originali da venire eletto alla Royal Society. Oltre che di matematica combinatoria, si occupò in modo proficuo di nodi, gruppi finiti e quaternioni. C'è una ben nota configurazione in geometria proiettiva, chiamata esagramma mistico di Pascal (sei punti su una curva conica uniti in tutti i modi possibili da segmenti di retta), in cui certe intersezioni sono ancora note come punti di Kirkman.

Disco per risolvere il problema delle studentesse di Kirkman, con n = 9.

Kirkman era noto per il suo pungente sarcasmo, che spesso esercitava nei confronti della filosofia di Herbert Spencer. Spesso si citava la sua parodia della definizione di evoluzione data da Spencer: «A change from a nohowish untalkaboutable all-likeness, to a somehowish and in-general-talk-aboutable not-all-likeness, by continuous somethingelseifications and sticktogetherations».

Kirkman pubblicò per la prima volta il suo problema delle studentesse nel 1847, in un giornale inglese di matematica, e il problema apparve nuovamente in *The Lady's and Gentleman's Diary for the Year 1850*. Qui è dato nella sua formulazione originale. Ogni giorno della settimana un insegnante porta a fare una passeggiata 15 studentesse. Durante la passeggiata le ragazze sono raggruppate in triple. L'insegnante può organizzare le triple in modo che, dopo le sette passeggiate, ogni coppia di ragazze sia stata nella stessa tripla una e una sola volta?

Qualsiasi soluzione di questo problema è naturalmente un sistema di triple di Steiner, ma delle 80 soluzione base per $n = 15$ solo sette sono soluzioni base per il problema delle studentesse. Costruzioni di Kirkman è il nome dato ai sistemi di triple di Steiner a cui sia stato anche imposto il requisito che le triple siano raggruppate in modo che ogni gruppo esaurisca tutti gli oggetti.

Il numero di coppie di ragazze è ancora $1/2n(n-1)$ e il numero di giorni richiesto per le passeggiate è $1/2\ (n-1)$. Questi valori sono interi solo quando n è un multiplo dispari di 3. Allora la sequenza di possibili valori è 3, 9, 15, 21 e così via, ossia la sequenza per i sistemi di triple di Steiner coi numeri presi uno sì e uno no. Ogni valore della sequenza ha una soluzione? Da quando Kirkman ha sollevato la questione sono state riempite centinaia di pagine sul problema, molte delle quali scritte da eminenti matematici. Il caso di $n = 3$ rimane banale. Le tre ragazze vanno semplicemente a fare una passeggiata. Il caso di nove ragazze in quattro giorni ha un'unica soluzione:

123	147	159	168
456	258	267	249
789	369	348	357

Come nei sistemi di triple di Steiner, i numeri di una tripla non sono ordinati e non ha importanza il modo in cui i numeri sono permutati, il modo in cui le triple sono disposte all'interno di ciascun gruppo o il modo in cui le cifre sono scambiate tra loro. Tutte le variazioni ottenute con queste permutazioni sono considerate appartenenti alla stessa soluzione.

Ci sono molti nuovi metodi, anche geometrici, per ottenere costruzioni di Kirkman. Uno di essi avrebbe deliziato Raimondo Lullo, il teologo spagnolo del XIII secolo la cui *Ars Magna* esplorava combinazioni di simboli con l'aiuto di dischi concentrici ruotanti. Per trovare una soluzione per $n = 9$ si disegni un cerchio e si scrivano sulla circonferenza, a distanze regolari, le cifre da 1 a 8. Si fissi poi al cerchio un disco di cartoncino della stessa grandezza, con uno spillo che passi per i due centri. Al centro del disco si scriva la cifra 9. Sul disco si disegni un diametro e due triangoli scaleni nel modo illustrato dalla figura della pagina a fronte.

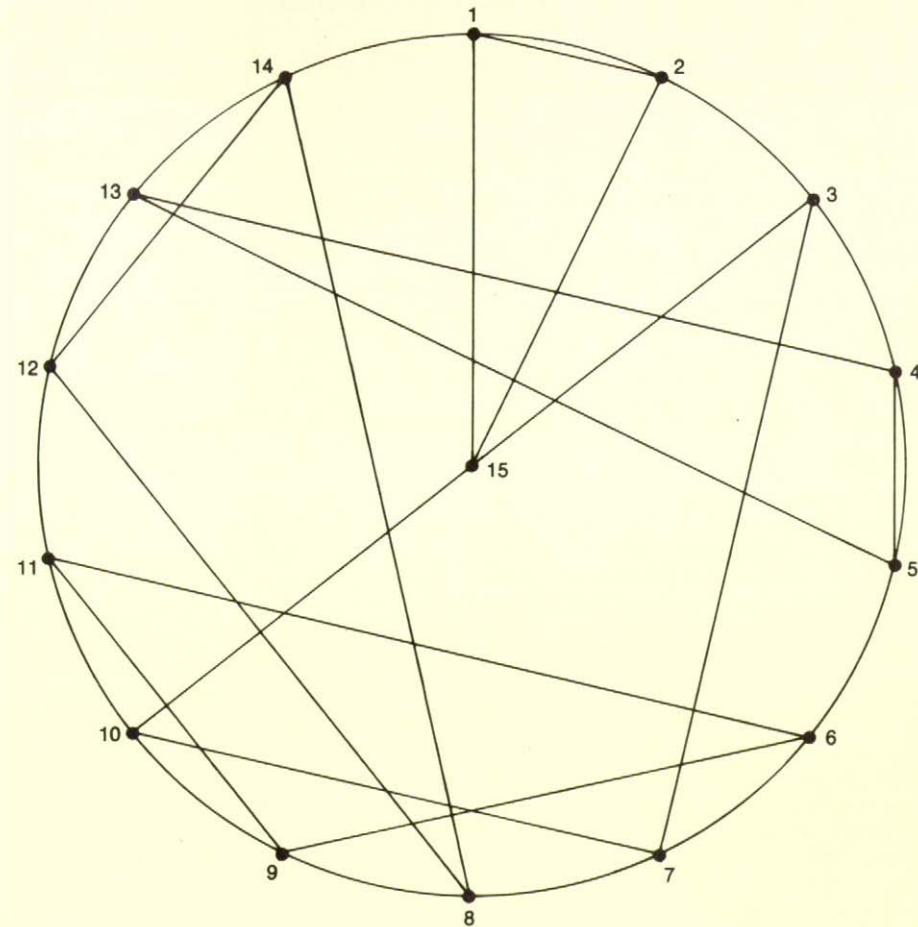
Si ruoti ora il cerchio, in una o l'altra direzione, un passo alla volta in quattro posizioni differenti. (Il quinto passo riporta la configurazione allo stato originario.) A ogni passo si copi la tripla indicata dalle estremità e dal centro della linea retta e le due triple indicate dagli angoli dei due triangoli. Le tre triple trovate in ciascuna delle quattro posizioni del disco danno le triple per ciascuno dei quattro giorni. Questa soluzione sembra differente dalla costruzione data in precedenza per il problema delle studentesse, ma sostituendo 2 a 5, 3 a 7, 4 a 9, 5 a 3, 6 a 8, 7 a 6, 8 a 4 e 9 a 2 (e lasciando inalterato 1) si ottiene la stessa costruzione. L'unico altro modo per sistemare dei triangoli sul disco in modo da generare una soluzione è disegnare l'immagine speculare della configurazione illustrata in figura. Questa procedura, però, non dà luogo a una altra costruzione.

Sin dal 1922 si sa che il caso di $n = 15$ ha sette soluzioni base. Esse possono essere generate da differenti configurazioni di triangoli, con o senza linea diametrale. Una configurazione di cinque triangoli è illustrata nella figura di questa pagina. In questo caso il disco deve essere ruotato di due unità alla volta in sette differenti posizioni. A ogni posizione gli angoli di ciascun triangolo danno una delle cinque triple per quel giorno.

Disco per risolvere il problema delle studentesse di Kirkman, con n = 15.

Va notato che nessuna coppia di triangoli su un disco può essere congruente. Se lo fossero, ci sarebbero delle triple doppie nella costruzione globale. Un lavoro classico sulle costruzioni di Kirkman è il capitolo 10 dell'undicesima edizione delle *Mathematical Recreations & Essays*, di W. W. Rouse Ball, nella revisione di H. S. M. Coxeter. Ottimo è anche lo stesso capitolo nella dodicesima edizione del libro (University of Toronto Press, 1974), completamente riscritto da Coxeter. Il nuovo capitolo sostituisce la storia delle costruzioni con una discussione sul modo in cui esse si collegano ad argomenti quali la geometria affine e proiettiva, le matrici di Hadamard, i codici di correzione d'errore, i quadrati latini.

Esiste una costruzione di Kirkman per ogni possibile valore di n? Sorprendentemente, questa domanda è rimasta senza risposta fino al 1970, quando D. K. Ray-Chaudhuri e Richard M. Wilson, dell'Ohio State University, hanno dimostrato che la risposta è sì. Il numero di soluzioni, però, rimane sconosciuto per valori di n maggiori di 20. La dimostrazione si trova in «Solution of Kirkman's Schoolgirl Problem», in *Combinatorics (Proceedings of Symposia in Pure Mathematics*, Vol. 19, pagine 187-203; 1971).

Le costruzioni di Kirkman hanno molti usi pratici. Ecco un modo tipico per applicare la costruzione $n = 9$ a un esperimento biologico. Supponiamo che un ricercatore voglia studiare l'effetto di nove condizioni ambientali su un certo animale. Ci sono quattro specie dell'animale e un singolo animale può subire ripercussioni differenti a seconda che sia giovane, del tutto sviluppato o vecchio. Ogni specie è assegnata a caso a uno di quattro gruppi. All'interno di ciascun gruppo ci sono tre triple, ciascuna delle quali include un animale di ciascuna categoria d'età scelto a caso. Ogni animale è ora assegnato a uno dei nove ambienti secondo lo schema dei nove numeri nel suo gruppo. Questa costruzione rende possibile un'analisi statistica estremamente semplice dei risultati per determinare quale effetto abbia l'ambiente indipendentemente dalle differenze d'età e specie.

Ho detto in precedenza che Kirkman introdusse una condizione aggiuntiva che trasformava i sistemi di triple di Steiner in un nuovo tipo di problema di block-design theory. Nel 1917, il genio dei rompicapo inglese Ernest Dudeney, impose una nuova restrizione alle costruzioni di Kirkman che dava luogo a un altro problema analogo (si veda il problema 272 del libro di *Dudeney Amusements in Mathematics* e il problema 287 del suo lavoro postumo *Puzzles and Curious Problems*).

«C'erano una volta, - inizia la seconda cornice narrativa del rompicapo di Dudeney - nove prigionieri di carattere particolarmente pericoloso che dovevano essere controllati con estrema attenzione. Ogni giorno della settimana erano portati fuori per fare degli esercizi e venivano ammanettati come si vede nello schizzo fatto da una delle loro guardie *(si veda la figura della pagina seguente)*. In nessun giorno della settimana gli stessi due uomini erano ammanettati insieme. Si può vedere come essi erano fatti uscire il lunedì. Sapete sistemare i nove uomini in triple per i rimanenti cinque giorni? Si vede che il numero 1 non può di nuovo essere ammanettato con il numero 2 (da una parte o dall'altra), né il numero 2 col numero 3, ma naturalmente il numero 1 e il numero 3 possono essere messi insieme. È perciò un problema ben differente dal vecchio problema delle quindici studentesse: ci si accorgerà che è un rompicapo affascinante e ripagherà del tempo passato alla ricerca della soluzione».

Dudeney dava una soluzione senza spiegare come arrivare a essa e ad altre soluzioni analoghe. Soluzioni che si possono comunque facilmente trovare applicando la tecnica di Lullo con due ruote. Una coppia campione è illustrata dalla figura in alto a pagina 120. Ogni disco viene ruotato, diciamo in senso orario, tre passi alla volta. A ogni passo viene generata una tripla dagli angoli dei tre triangoli. In questo caso ogni tripla deve avere al centro il numero indicato dall'angolo che racchiude un punto.

Ogni disco genera i tre gruppi elencati sotto. In entrambi gli insiemi i gruppi sono ciclici nel senso che se si aggiunge 3 (modulo 7) a ogni numero del primo gruppo si ottiene il secondo gruppo. Analogamente, il secondo gruppo genera il terzo, il quale a sua volta riporta al primo. La soluzione non inizia con la combinazione data da Dudeney per il primo giorno, anche se è facile arrivarci scambiando le cifre.

Dopo aver dato la risposta, Dudeney continuava in tono di sfida: «Se il lettore vuole un rompicapo difficile per rimanere occupato durante i mesi invernali, provi a disporre ventun prigionieri in modo che possano tutti uscire per quindici giorni, ammanettati in triple, senza che due uomini siano ammanettati insieme più di una volta. Nel caso si convinca che il compito è impossibile, aggiungeremo che abbiamo ottenuto una soluzione perfetta. Ma è un'ardua impresa!».

È davvero un'ardua impresa. A quanto ne so, la prima soluzione pubblicata si trova in *Graph Decompositions, Handcuffed Prisoners and Balanced P-Designs*, di Pavol Hell e Alexander Rosa, in «Discrete Mathematics» (Vol. 2, n. 3. pagine 229-252; giugno 1972).

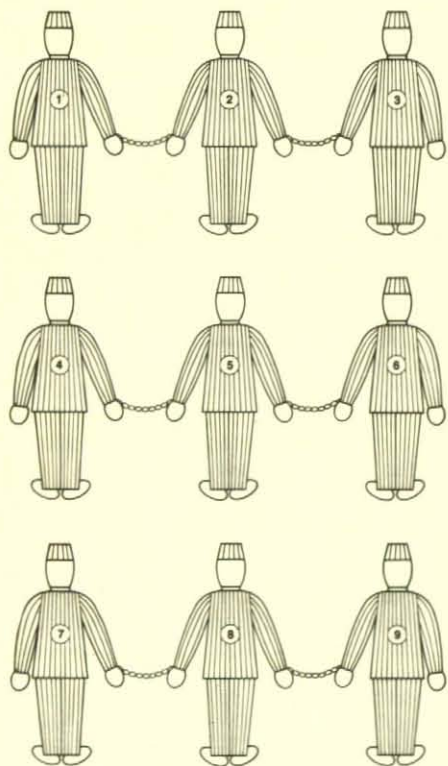
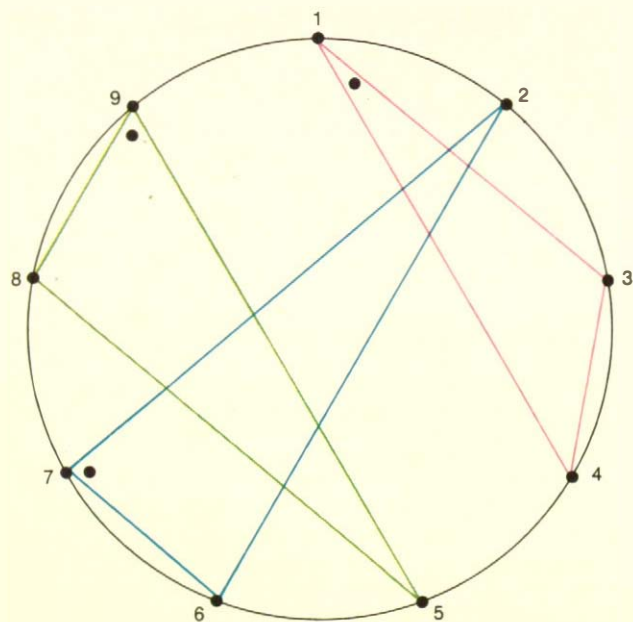
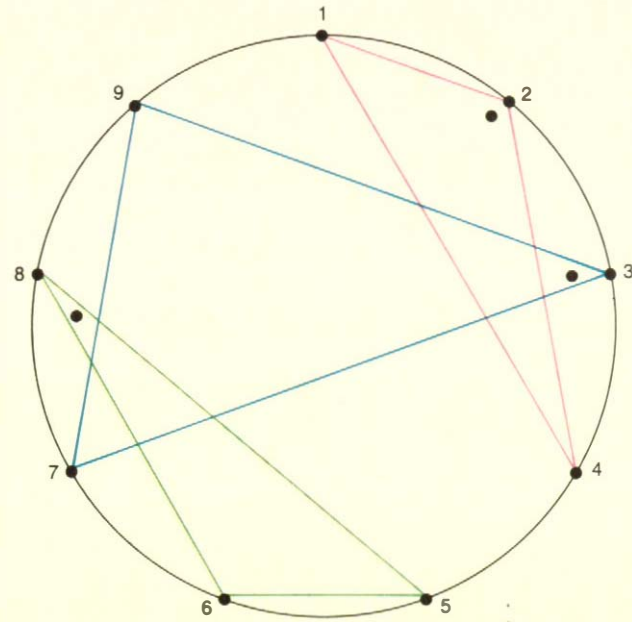
Prima di dare la soluzione, vorrei fare qualche osservazione generale sul problema dei prigionieri ammanettati. Il numero di coppie di prigionieri è $1/2n(n-1)$, come avveniva con i sistemi di triple di Steiner e le costruzioni di Kirkman, anche se la nuova restrizione (le manette!) allunga il numero di giorni richiesto a $3/4\ (n-1)$. C'è una soluzione solo quando questa espressione è un intero, il che avviene quando n ha un valore compreso in una sequenza formata da esattamente la metà dei valori possibili per una costruzione di Kirkman, vale a dire la sequenza 9, 21, 33, 45, 57, 69, 81, 93 e così via, in cui la differenza tra ogni coppia adiacente di interi è 12.

Problema dei prigionieri ammanettati, n = 9.

Nel 1971, Charlotte Huang e Rosa pubblicarono una classificazione di 334 soluzioni base per $n = 9$. Quando però Dame Kathleen Ollerenshaw e il cosmologo Hermann Bondi analizzarono ciascuna soluzione, trovarono due duplicati tra le 334. Si ritiene ora che l'effettivo numero di soluzioni sia 332. Per tutti i valori di n maggiori di 9 il numero di soluzioni non è noto. Per $n = 21$, Rosa ritiene che esistano milioni di soluzioni. Hall e Rosa hanno dimostrato che un numero infinito di n ammettono soluzioni e hanno indicato il modo per trovare soluzioni cicliche per tutti gli n minori di 100, con l'eccezione di 57, 69 e 93. Wilson (che aveva contribuito a risolvere il problema delle studentesse di Kirkman) ha dimostrato con un lavoro non ancora pubblicato che tutti i valori di n ammettono una soluzione.

Nella figura in basso di pagina 120 si vede una soluzione ciclica trovata da Hell e Rosa per $n = 21$. I primi sette giorni formano un insieme ciclico che può essere generato da un disco con sette triangoli i cui angoli corrispondono alle triple che sono in testa alla costruzione di ciascun giorno. Il disco viene ruotato di tre passi alla volta. Un secondo disco con sette triangoli genera analogamente i successivi sette giorni, e il quindicesimo giorno ha la costruzione che si vede a destra nella figura. In entrambi gli insiemi ciclici, la costruzione di un giorno può essere scambiata con quella del giorno successivo aggiungendo 3 (modulo 21) a ciascun numero; la stessa operazione applicata all'ultimo giorno riporta lo schema al primo giorno. Hell e Rosa forniscono soluzioni cicliche analoghe per $n = 33$ e $n = 45$.

Sia il problema delle studentesse sia quello dei prigionieri possono essere generalizzati a quadruple, quintuple, sestuple e così via. Tale generalizzazione porta a profondi enigmi combinatori, molti dei quali sono lungi dall'essere risolti. Sui libri di rompicapo si trovano centinaia di problemi simili, spesso inseriti in storielle su disposizioni di posti a sedere, tornei, membri di associazioni o altri schemi combinatori. Per esempio, mi viene chiesto spesso come si possono disporre n membri di un club di bridge (n deve essere un multiplo di 4) perché si possano incontrare giornalmente per $n-1$ giorni in modo che ogni giocatore sia il compagno di ciascun altro giocatore esattamente una volta e avversario di ciascun altro giocatore esattamente due volte.

Il problema sembra piuttosto semplice, ma in realtà è così spinoso da non essere stato risolto che pochi anni fa. La migliore analisi fattane si può trovare in *Whist Tournaments*, un articolo scritto da Ronald D. Baker, dell'Università del Delaware. (L'articolo uscì in «Proceedings of the Sixth Southeastern Conference on Combinatorics, Graph Theory and Computing», pubblicati nel 1975 da Utilitas

	I GIORNO			II GIORNO			III GIORNO		
ROSSO	4	1	3	7	4	6	1	7	9
BLU	2	7	6	5	1	9	8	4	3
VERDE	5	9	8	8	3	2	2	6	5

	IV GIORNO			V GIORNO			VI GIORNO		
ROSSO	1	2	4	4	5	7	7	8	1
BLU	7	3	9	1	6	3	4	9	6
VERDE	5	8	6	8	2	9	2	5	3

Una soluzione per i prigionieri ammanettati, n = 9.

Mathematica, Winnipeg, come Volume 14 della serie *Congressus Numerantium.*) Baker mostra come trovare soluzioni per tutti i valori di n tranne 132, 152 e 264. In seguito il caso di $n = 132$ è stato risolto dal matematico israeliano Hain Hanani e quelli di $n = 152$ e $n = 264$ da Baker e Wilson, i quali pubblicheranno presto un'analisi dettagliata dei tre casi.

Per molti valori di n, si possono generare soluzioni con dischi che ruotano di un passo alla volta. Nella figura della pagina a fronte si possono vedere dischi per $n = 4$ e $n = 8$. La tecnica per generare soluzioni è semplice. Si traccia una linea da 1 (il centro del disco) a 2. Si traccia poi un'altra linea per unire altri due numeri. I punti terminali di ciascuna linea sono compagni al tavolo di bridge e le coppie di compagni sono avversarie allo stesso tavolo. Se c'è un altro tavolo, si uniscono altre due coppie di numeri con linee colorate che indicano la disposizione dei posti a sedere per il tavolo. Per ulteriori tavoli si aggiungono altre linee.

La disposizione di queste linee genera una soluzione ciclica se e solo se vengono rispettate due condizioni. Primo, non ci sono due linee della stessa lunghezza (in quanto la lunghezza è data dal numero di unità che una linea congiunge sulla circonferenza). Tranne nel caso del raggio le lunghezze saranno necessariamente interi consecutivi da 1 a $1/2n$-1. Secondo, se tutti gli avversari a ciascun tavolo sono uniti da linee (tratteggiate nella figura), ogni lunghezza comparirà sul disco solo due volte.

Le linee sono per lo più sistemate andando per tentativi. Non è nota alcuna procedura che garantisca uno schema corretto per tutti i valori di n. Una volta trovato lo schema, esso indica la disposizione dei posti a sedere per il primo giorno. Ruotando il disco si generano le disposizioni per gli altri giorni. Ogni colonna della costruzione finale è ciclica, così che una volta determinata la disposizione dei posti per il primo giorno, si può rapidamente determinare una tabella dei posti per i rimanenti giorni senza dover ruotare il disco. Le soluzioni inedite per $n = 132$, $n = 152$ e $n = 264$ non sono cicliche, anche se forse è possibile metterle in forma ciclica permutando i numeri. Secondo Baker, tutti i valori di n possono avere soluzioni cicliche, sebbene non sia noto alcun algoritmo per trovarle.

Ecco ora un simpatico problema a cui darò risposta il mese prossimo: riuscite a costruire un disco per 12 giocatori di bridge che generi un torneo ciclico rispondente alle condizioni desiderate?

1	2	3	4	5	6	7
1 8 18	4 11 21	7 14 3	10 17 6	13 20 9	16 2 12	19 5 15
2 4 20	5 7 2	8 10 5	11 13 8	14 16 11	17 19 14	20 1 17
3 7 15	6 10 18	9 13 21	12 16 3	15 19 6	18 1 9	21 4 12
10 11 6	13 14 9	16 17 12	19 20 15	1 2 18	4 5 21	7 8 3
5 16 21	8 19 3	11 1 6	14 4 9	17 7 12	20 10 15	2 13 18
19 9 17	1 12 20	4 15 2	7 18 5	10 21 8	13 3 11	16 6 14
13 12 14	16 15 17	19 18 20	1 21 2	4 3 5	7 6 8	10 9 11

8	9	10	11	12	13	14	15
1 4 19	4 7 1	7 10 4	10 13 7	13 16 10	16 19 13	19 1 16	1 3 2
7 16 9	10 19 12	13 1 15	16 4 18	19 7 21	1 10 3	4 13 6	4 6 5
10 2 6	13 5 9	16 8 12	19 11 15	1 14 18	4 17 21	7 20 3	7 9 8
13 17 8	16 20 11	19 2 14	1 5 17	4 8 20	7 11 2	10 14 5	10 12 11
11 14 20	14 17 2	17 20 5	20 2 8	2 5 11	5 8 14	8 11 17	13 15 14
5 12 21	8 15 3	11 18 6	14 21 9	17 3 12	20 8 15	2 9 18	16 18 17
3 18 15	6 21 18	9 3 21	12 6 3	15 9 6	18 12 9	21 15 12	19 21 20

Una soluzione per i prigionieri ammanettati, n = 21.

Nel problema del treno dato nella rubrica di aprile non avrei dovuto far fermare il treno lungo il tragitto di 500 miglia. Henry S. Warren, Jr., è stato il primo di molti lettori a far notare che, se vengono effettuate fermate, il tempo impiegato dal treno per percorrere un segmento di 50 miglia non è una funzione continua della posizione del segmento nel suo muoversi da un'estremità all'altra del tragitto.

Se invece di far scivolare un segmento di 50 miglia, si fa scivolare un segmento di un'ora nella direzione del tempo lungo un grafo spazio-tempo, allora si può davvero dimostrare che ci dev'essere almeno un'ora in cui il treno si è spostato di sole 50 miglia anche se sono consentite fermate e spostamenti all'indietro. Non è possibile affermare, invece, che c'è una tratta di 50 miglia percorsa in un'ora esatta.

Per chiarire questo punto, supponiamo che il treno abbia viaggiato costantemente a 100 miglia all'ora tranne che per una fermata di cinque ore a metà giornata. Dopo aver fatto le ultime 50 miglia in 30 minuti prima di fermarsi, il macchinista potrebbe guardare l'orologio e dopo 30 minuti, con il treno ancora fermo, potrebbe veridicamente affermare che il treno ha fatto 50 miglia nell'ultima ora. Eppure, solo in senso pickwickiano si può dire che il treno abbia fatto quelle 50 miglia a una media di 50 miglia all'ora, perché appena un secondo dopo la media per lo stesso segmento è ridotta e decresce finché il treno non riparte.

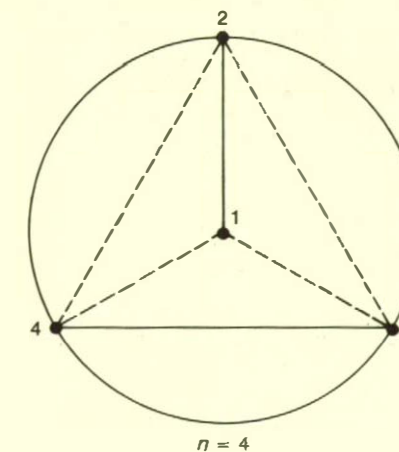
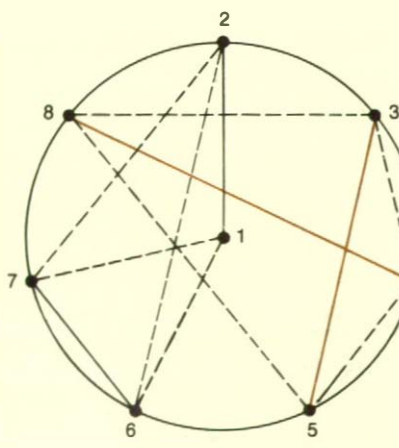
Nel mio articolo di giugno avevo promesso di riprendere in considerazione lo sfortunato «problema impossibile». Molti lettori hanno costruito programmi di calcolo che mostrano come, portando sopra 100 il limite superiore per i due numeri, appaiono continuamente nuove soluzioni. La coppia di numeri successiva è 4 e 61. Non è vero, quindi, che 4 e 13 sia una risposta unica indipendentemente dall'altezza del limite.

Un programma elaborato da David Pearson, del Dortmouth College, identificava 78 coppie che risolvono il problema quando la somma dei due numeri è minore di 2000 e il limite è sufficientemente alto. In tutti i casi il più piccolo dei due numeri è una potenza di 2 e in tutti tranne cinque casi il numero più grande è primo. Si ritiene che il numero più piccolo sia sempre una potenza di 2, ma non mi risulta che ci sia nessuna prova. Non è dimostrato, inoltre, che senza limite superiore il numero di coppie sia infinito.

Un lettore di Amsterdam, Robert W. van der Waal, mi ha mandato una copia del problema nella presentazione fattane sulla rivista olandese di matematica «Nieuw Archief Voor Wiskunde» (Serie 3, Vol. 17, pagina 152, 1969, e Vol. 18, pagine 102-106, 1970). Non conosco alcuna indicazione bibliografica anteriore.

Nella mia rubrica del giugno scorso ho analizzato il problema di sistemare n super-regine (una super-regina è un pezzo degli scacchi che unisce i movimenti di una regina e di un cavallo) su una scacchiera n^2 in modo che non ci siano due pezzi uno sotto l'attacco dell'altro. Per valori di n maggiori di 9 è stato dimostrato che devono esistere soluzioni per tutti gli n che non siano della forma $(12k+8)$ o $(12k+9)$, dove k è un intero. In altre parole, manca la dimostrazione per valori appartenenti alla sequenza 20, 21, 32, 33, 34, 45, 56, 57 e così via. Un gran numero di lettori, troppi per essere menzionati, hanno inviato soluzioni per i primi sei di questi valori. J. Reineke e P. Päppinghaus, dell'Università di Hannover, hanno mandato soluzioni fino a $n = 69$, lasciando 80 come minimo caso irrisolto.

Molti lettori della mia rubrica di maggio sulla dama hanno mandato una seconda soluzione al problema di come una dama nera possa pattare contro due dame bianche in una posizione in cui il Bianco muove per primo. Avrei dovuto aggiungere «e non perde immediatamente una dama». Senza questa condizione ci sono due soluzioni banali oltre a quella data in giugno. In una soluzione il Bianco si trova su 10 e 19 e il Nero su 15 tra le due dame. Nell'altra soluzione il Bianco occupa, diciamo, il 30 e il 21 e il Nero il 22.

John Harris ha studiato il problema consistente nello scambiare le posizioni del Bianco e del Nero nel minimo numero di mosse seguendo le regole del gioco. Ha abbassato il record che avevo dato da 172 mosse a 160. Harris, tra l'altro, detiene anche il primato per l'altro compito da me proposto: scambiare le posizioni del Bianco e del Nero in modo che alla fine tutti e 24 i pezzi siano dame. Il suo record di 180 mosse (90 per lato) è imbattuto.

GIORNI	TAVOLO I	
1	12	34
2	13	42
3	14	23

GIORNI	TAVOLO I		TAVOLO 2	
1	12	67	35	48
2	13	78	46	52
3	14	82	57	63
4	15	23	68	74
5	16	34	72	85
6	17	45	83	26
7	18	56	24	37

Costruzioni per il torneo di bridge con n = 4 e n = 8.